

BEST AVAILABLE COPY

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau



AB

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12N 15/12, 15/70, 15/75, 15/74, 15/76, C07K 13/00, C12N 1/21 // C12R 1:125, (C12N 1/21, C12R 1:19)		A2	(11) International Publication Number: WO 94/29450 (43) International Publication Date: 22 December 1994 (22.12.94)																																																																											
(21) International Application Number: PCT/US94/06689 (22) International Filing Date: 15 June 1994 (15.06.94)		(81) Designated States: CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).																																																																												
(30) Priority Data: 08/077,600 15 June 1993 (15.06.93) US		Published <i>Without international search report and to be republished upon receipt of that report.</i>																																																																												
(60) Parent Application or Grant (63) Related by Continuation US 08/077,600 (CIP) Filed on 15 June 1993 (15.06.93)																																																																														
(71) Applicant (for all designated States except US): E.I. DU PONT DE NEMOURS AND COMPANY [US/US]; 1007 Market Street, Wilmington, DE 19898 (US).																																																																														
(72) Inventor; and (75) Inventor/Applicant (for US only): FAHNESTOCK, Stephen, R. [US/US]; 719 Mt. Lebanon Road, Wilmington, DE 19803-1609 (US).																																																																														
(74) Agents: FLOYD, Linda, Axamethy et al.; E.I. du Pont de Nemours and Company, Legal/Patent Records Center, 1007 Market Street, Wilmington, DE 19898 (US).																																																																														
(54) Title: NOVEL, RECOMBINANTLY PRODUCED SPIDER SILK ANALOGS																																																																														
(57) Abstract																																																																														
<p>The invention relates to novel spider silk protein analogs derived from the amino acid consensus sequence of repeating units found in the natural spider dragline of <i>Nephila clavipes</i>. More specifically, synthetic spider dragline has been produced from <i>E. coli</i> and <i>Bacillus subtilis</i> recombinant expression systems wherein expression from <i>E. coli</i> is at levels greater than 1 mg full-length polypeptide per gram of cell mass.</p> <table border="0"> <tr> <td>1</td> <td>...</td> <td>QG A GAAAAAA-GG</td> </tr> <tr> <td>2</td> <td>A GQG GYG GLG GQG</td> <td>--- --- - - - - -</td> </tr> <tr> <td>3</td> <td>A GQG GYG GLG GQG A</td> <td>--- --- GQG A GAAAAAAAGG</td> </tr> <tr> <td>4</td> <td>A GQG GYG GLG SQG A GRG</td> <td>--- GQG A GAAAAAA-GG</td> </tr> <tr> <td>5</td> <td>A GQG GYG GLG SQG A GRG GLG</td> <td>GQG A GAAAAAAAGG</td> </tr> <tr> <td>6</td> <td>A GQG GYG GLG NQG A GRG</td> <td>--- GQG - - - - -</td> </tr> <tr> <td>7</td> <td>A GQG GYG GLG SQG A GRG GLG</td> <td>GQG A GAAAAAA-GG</td> </tr> <tr> <td>8</td> <td>A GQG GYG GLG GQG</td> <td>--- --- - - - - -</td> </tr> <tr> <td>9</td> <td>A GQG GYG GLG SQG A GRG GLG</td> <td>GQG A GAAAAAAAGG</td> </tr> <tr> <td>10</td> <td>A GQG --- GLG GQG A</td> <td>--- GQG A GASAAA-GG</td> </tr> <tr> <td>11</td> <td>A GQG GYG GLG SQG A GRG</td> <td>--- GEG A GAAAAAA-GG</td> </tr> <tr> <td>12</td> <td>A GQG GYG GLG GQG</td> <td>--- --- - - - - -</td> </tr> <tr> <td>13</td> <td>A GQG GYG GLG SQG A GRG GLG</td> <td>GQG A GAAA---GG</td> </tr> <tr> <td>14</td> <td>A GQG --- GLG GQG A</td> <td>--- GQG A GAAAAAA-GG</td> </tr> <tr> <td>15</td> <td>A GQG GYG GLG SQG A GRG GLG</td> <td>GQG A GAVAAAAAGG</td> </tr> <tr> <td>16</td> <td>A GQG GYG GLG SQG A GRG</td> <td>--- GQG A GAAAAAA-GG</td> </tr> <tr> <td>17</td> <td>A GQR GYG GLG NQG A GRG GLG</td> <td>GQG A GAAAAAAAGG</td> </tr> <tr> <td>18</td> <td>A GQG GYG GLG NQG A GRG</td> <td>--- GQG - - - - -</td> </tr> <tr> <td>19</td> <td>A GQG GYG GLG SQG A GRG</td> <td>--- GQG A GAAAAAA-VG</td> </tr> <tr> <td>20</td> <td>A GQE --- GIR GQG</td> <td>--- --- - - - - -</td> </tr> <tr> <td>21</td> <td>A GQG GYG GLG SQG S GRG</td> <td>GLG GQG A GAAAAAA-GG</td> </tr> <tr> <td>22</td> <td>A GQG --- GLG GQG A</td> <td>--- GQG A GAAAAAA-GG</td> </tr> <tr> <td>23</td> <td>V RQG GYG GLG SQG A GRG</td> <td>--- GQG A GAAAAAA-GG</td> </tr> <tr> <td>24</td> <td>A GQG GYG GLG GQG V GRG</td> <td>GLG GQG A GAAA---GG</td> </tr> <tr> <td>25</td> <td>A GQG GYG GVG S</td> <td>--- --- - - - - - G A SAAAAGG</td> </tr> </table>				1	...	QG A GAAAAAA-GG	2	A GQG GYG GLG GQG	--- --- - - - - -	3	A GQG GYG GLG GQG A	--- --- GQG A GAAAAAAAGG	4	A GQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-GG	5	A GQG GYG GLG SQG A GRG GLG	GQG A GAAAAAAAGG	6	A GQG GYG GLG NQG A GRG	--- GQG - - - - -	7	A GQG GYG GLG SQG A GRG GLG	GQG A GAAAAAA-GG	8	A GQG GYG GLG GQG	--- --- - - - - -	9	A GQG GYG GLG SQG A GRG GLG	GQG A GAAAAAAAGG	10	A GQG --- GLG GQG A	--- GQG A GASAAA-GG	11	A GQG GYG GLG SQG A GRG	--- GEG A GAAAAAA-GG	12	A GQG GYG GLG GQG	--- --- - - - - -	13	A GQG GYG GLG SQG A GRG GLG	GQG A GAAA---GG	14	A GQG --- GLG GQG A	--- GQG A GAAAAAA-GG	15	A GQG GYG GLG SQG A GRG GLG	GQG A GAVAAAAAGG	16	A GQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-GG	17	A GQR GYG GLG NQG A GRG GLG	GQG A GAAAAAAAGG	18	A GQG GYG GLG NQG A GRG	--- GQG - - - - -	19	A GQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-VG	20	A GQE --- GIR GQG	--- --- - - - - -	21	A GQG GYG GLG SQG S GRG	GLG GQG A GAAAAAA-GG	22	A GQG --- GLG GQG A	--- GQG A GAAAAAA-GG	23	V RQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-GG	24	A GQG GYG GLG GQG V GRG	GLG GQG A GAAA---GG	25	A GQG GYG GVG S	--- --- - - - - - G A SAAAAGG
1	...	QG A GAAAAAA-GG																																																																												
2	A GQG GYG GLG GQG	--- --- - - - - -																																																																												
3	A GQG GYG GLG GQG A	--- --- GQG A GAAAAAAAGG																																																																												
4	A GQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-GG																																																																												
5	A GQG GYG GLG SQG A GRG GLG	GQG A GAAAAAAAGG																																																																												
6	A GQG GYG GLG NQG A GRG	--- GQG - - - - -																																																																												
7	A GQG GYG GLG SQG A GRG GLG	GQG A GAAAAAA-GG																																																																												
8	A GQG GYG GLG GQG	--- --- - - - - -																																																																												
9	A GQG GYG GLG SQG A GRG GLG	GQG A GAAAAAAAGG																																																																												
10	A GQG --- GLG GQG A	--- GQG A GASAAA-GG																																																																												
11	A GQG GYG GLG SQG A GRG	--- GEG A GAAAAAA-GG																																																																												
12	A GQG GYG GLG GQG	--- --- - - - - -																																																																												
13	A GQG GYG GLG SQG A GRG GLG	GQG A GAAA---GG																																																																												
14	A GQG --- GLG GQG A	--- GQG A GAAAAAA-GG																																																																												
15	A GQG GYG GLG SQG A GRG GLG	GQG A GAVAAAAAGG																																																																												
16	A GQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-GG																																																																												
17	A GQR GYG GLG NQG A GRG GLG	GQG A GAAAAAAAGG																																																																												
18	A GQG GYG GLG NQG A GRG	--- GQG - - - - -																																																																												
19	A GQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-VG																																																																												
20	A GQE --- GIR GQG	--- --- - - - - -																																																																												
21	A GQG GYG GLG SQG S GRG	GLG GQG A GAAAAAA-GG																																																																												
22	A GQG --- GLG GQG A	--- GQG A GAAAAAA-GG																																																																												
23	V RQG GYG GLG SQG A GRG	--- GQG A GAAAAAA-GG																																																																												
24	A GQG GYG GLG GQG V GRG	GLG GQG A GAAA---GG																																																																												
25	A GQG GYG GVG S	--- --- - - - - - G A SAAAAGG																																																																												

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

TITLE

NOVEL, RECOMBINANTLY PRODUCED SPIDER SILK ANALOGS

FIELD OF THE INVENTION

The invention relates to novel spider silk protein
5 analogs derived from the amino acid consensus sequence
of repeating units found in the natural spider dragline
of *Nephila clavipes*. More specifically, synthetic
spider dragline has been produced from *E. coli* and
10 *Bacillus subtilis* recombinant expression systems wherein
expression from *E. coli* is at levels greater than 1 mg
full-length polypeptide per gram of cell mass.

BACKGROUND

Ever increasing demands for materials and fabrics
that are both light-weight and flexible without
15 compromising strength and durability has created a need
for new fibers possessing higher tolerances for such
properties as elasticity, denier, tensile strength and
modulus. The search for a better fiber has led to the
investigation of fibers produced in nature, some of
20 which possess remarkable qualities. The virtues of
natural silk produced by *Bombyx mori* (silk worm) have
been well known for years but it is only recently that
other other naturally produced silks have been examined.

Spider silks have been demonstrated to have several
25 desirable characteristics. The orb-web-spinning spiders
can produce silk from six different types of glands.
Each of the six fibers has different mechanical
properties. However, they all have several features in
common. They are (i) composed predominantly or
30 completely of protein; (ii) undergo a transition from a
soluble to an insoluble form that is virtually
irreversible; (iii) composed of amino acids dominated by
alanine, serine, and glycine and have substantial
quantities of other amino acids, such as glutamine,
35 tyrosine, leucine, and valine. The spider dragline silk

fiber has been proposed to consist of pseudocrystalline regions of antiparallel, β -sheet structure interspersed with elastic amorphous segments.

The spider silks range from those displaying a
5 tensile strength greater than steel (7.8 vs
3.4 G/denier) and those with an elasticity greater than
wool, to others characterized by energy-to-break limits
that are greater than KEVLAR[®] (1x10⁵ vs 3x10⁴ JKG-1).
Given these characteristics spider silk could be used as
10 a light-weight, high strength fiber for various textile
applications.

Considerable difficulty has been encountered in
attempting to solubilize and purify natural spider silk
while retaining the molecular-weight integrity of the
15 fiber. The silk fibers are insoluble except in very
harsh agents such as LiSCN, LiClO₄, or 88% (vol/vol)
formic acid. Once dissolved, the protein precipitates
if dialyzed or if diluted with typical buffers. Another
disadvantage of spider silk protein is that only small
20 amounts are available from cultivated spiders, making
commercially useful quantities of silk protein
unattainable at a reasonable cost. Additionally,
multiple forms of spider silks are produced
simultaneously by any given spider. The resulting
25 mixture has less application than a single isolated silk
because the different spider-silk proteins have
different properties and, due to solubilization
problems, are not easily separated by methods based on
their physical characteristics. Hence the prospect of
30 producing commercial quantities of spider silk from
natural sources is not a practical one and there remains
a need for an alternate mode of production. The
technology of recombinant genetics provides one such
mode.

By the use of recombinant DNA technology it is now possible to transfer DNA between different organisms for the purposes of expressing desired proteins in commercially useful quantities. Such transfer usually 5 involves joining appropriate fragments of DNA to a vector molecule, which is then introduced into a recipient organism by transformation. Transformants are selected by a known marker on the vector, or by a genetic or biochemical screen to identify the cloned 10 fragment. Vectors contain sequences that enable autonomous replication within the host cell, or allow integration into a chromosome in the host.

If the cloned DNA sequence encodes a protein, a series of events must occur to obtain synthesis of this 15 foreign protein in an active form in the host cell.

Promoter sequences must be present to allow transcription of the gene by RNA polymerase, and a ribosome binding site and initiation codon must be present in the transcribed mRNA for translation by 20 ribosomes. These transcriptional and translational recognition sequences are usually optimized for effective binding by the host RNA polymerase and ribosomes, and by the judicious choice of vectors, it is often possible to obtain effective expression of many 25 foreign genes in a host cell.

While many of the problems of efficient transcription and translation have been generally recognized and for the most part, overcome, the synthesis of fiber-forming foreign polypeptides 30 containing high numbers of repeating units poses unique problems. Genes encoding proteins of this type are prone to genetic instability due to the repeating nucleic acid sequences. Ideally, they encode proteins of high molecular weight, consisting of at least 800 35 amino acid residues, and generally with restricted amino

acid compositions. While *E. coli* produces endogenous proteins in excess of 1000 residues, production of long proteins of restricted amino acid composition appears to place an unbalanced strain on the biosynthetic system, 5 resulting in the production of truncated products, probably due to abortive translation.

In spite of the above mentioned difficulties, recombinant expression of fiber forming proteins is known in the art. Chatellard et al., *Gene*, 81, 267, 10 (1989) teach the cloning and expression of the trimeric fiber protein of human adenovirus type 2 from *E. coli*. The gene expression system relied upon bacteriophage T7 RNA polymerase and optimal gene expression was obtained at 30 °C where the foreign protein attained levels of 1% 15 of total host protein.

Goldberg et al., *Gene*, 80, 305, (1989) disclose the cloning and expression in *E. coli* of a synthetic gene encoding a collagen analog (poly (Gly-Pro-Pro)). The largest DNA insert was on the order of 450 base pairs 20 and it was suggested that large segments of highly-repeated DNA may be unstable in *E. coli*.

Ferrari et al. (WO 8803533) disclose methods and compositions for the production of polypeptides having repetitive oligomeric units such as those found in silk-like proteins and elastin-like proteins by the expression of synthetic structural genes. The DNA sequences of Ferrari encode peptides containing an oligopeptide repeating unit which contains at least 3 different amino acids and a total of 4-30 amino acids, 25 there being at least 2 repeating units in the peptide and at least 2 identical amino acids in each repeating unit.

Cappello et al. (WO 9005177) teach the production of a proteinaceous polymer from transformed prokaryotic 30 hosts comprising strands of repeating units which can be

assembled into aligned strands and DNA sequences encoding the same. The repeating units are derived from natural polymers such as fibroin, elastin, keratin or collagen.

5 The cloning and expression of silk-like proteins is also known. Ohshima et al., *Proc. Natl. Acad. Sci. U.S.A.*, 74, 5363, (1977) reported the cloning of the silk fibroin gene complete with flanking sequences of *Bombyx mori* into *E. coli*. Petty-Saphon et al. 10 (EP 230702) disclose the recombinant production of silk fibroin and silk sericin from a variety of hosts including *E. coli*, *Saccharomyces cerevisiae*, *Pseudomonas* sp *Rhodopseudomonas* sp, *Bacillus* sp, and *Streptomyces* sp. In the preferred embodiments the expression of silk 15 proteins derived from *Bombyx mori* is discussed.

Progress has also been made in the cloning and expression of spider silk proteins. Xu et al., *Proc. Natl. Acad. Sci. U.S.A.*, 87, 7120, (1990) report the determination of the sequence for a portion of the 20 repetitive sequence of a dragline silk protein, Spidroin 1, from the spider *Nephila clavipes*, based on a partial cDNA clone. The repeating unit is a maximum of 34 amino acids long and is not rigidly conserved. The repeat unit is composed of two different segments: (i) a 10 25 amino acid segment dominated by a polyalanine sequence of 5-7 residues; (ii) a 24 amino acid segment that is conserved in sequence but has deletions of multiples of 3 amino acids in many of the repeats. The latter sequence consists predominantly of GlyXaaGly motifs, 30 with Xaa being alanine, tyrosine, leucine, or glutamine. The codon usage for this DNA is highly selective, avoiding the use of cytosine or guanine in the third position.

Hinman and Lewis, *J. Biol. Chem.* 267, 19320 (1992) 35 report the sequence of a partial cDNA clone encoding a

portion of the repeating sequence of a second fibroin protein, Spidroin 2, from dragline silk of *Nephila clavipes*. The repeating unit of Spidroin 2 is a maximum of 51 amino acids long and is also not rigidly 5 conserved. The frequency of codon usage of the Spidroin 2 cDNA is very similar to Spidroin 1.

Lewis et al. (EP 452925) disclose the expression of spider silk proteins including protein fragments and variants, of *Nephila clavipes* from transformed *E. coli*. 10 Two distinct proteins were independently identified and cloned and were distinguished as silk protein 1 ((Spidroin 1) and silk protein 2 (Spidroin 2)).

Lombardi et al. (WO 9116351) teach the production of recombinant spider silk protein comprising an 15 amorphous domain or subunit and a crystalline domain or subunit where the domain or subunit refers to a portion of the protein containing a repeating amino acid sequence that provides a particular mechanostructural property.

20 The above mentioned expression systems are useful for the production of recombinant silks and silk variants, however all rely on the specific cloned gene of a silk producing organism. One detrimental effect of such systems is that codon usage is not optimized for 25 the production of foreign proteins in a recombinant host. It is well known in the art that expression of a foreign gene is more efficient if codons not favored by the organism in which expression is desired are avoided. Foreign genes cloned into recombinant hosts often rely 30 on a codon usage not typically found in the host. This often results in poor yields of foreign protein.

35 There remains a need therefore for a method to produce a spider silk protein in commercially useful quantities. It is the object of the present invention to meet such need by providing novel DNA sequences

encoding variants of consensus sequences derived from spider silk proteins capable of being expressed in a foreign host having the ability to produce synthetic proteins in commercially useful amounts of 1% to 30% of 5 total host protein.

SUMMARY OF THE INVENTION

The present invention provides novel synthetic spider dragline variant proteins produced by a process comprising the steps of: designing a DNA monomer sequence 10 of between about 50 bp and 1000 bp which codes for an polypeptide monomer consisting of a variant of a consensus sequence derived from the fiber forming regions of spider dragline protein; assembling the DNA monomer; polymerizing the DNA monomer to form a 15 synthetic gene encoding a full length silk variant protein; transforming a suitable host cell with a vector containing the synthetic gene; expressing the DNA polymer whereby the protein encoded by the DNA polymer is produced at levels greater than 1 mg full-length 20 protein per gram of cell mass and; recovering the protein in a useful form.

The present invention provides novel plasmids containing DNA compositions encoding spider silk variant proteins and novel transformed host cells containing 25 these plasmids which are capable of expressing the silk variant protein at levels greater than 1 mg full-length polypeptide per gram of cell mass.

Also included in the scope of the invention are transformed host cells capable of secreting full-length 30 spider dragline protein analogs into the cell growth medium.

In a preferred embodiment, an artificial gene is constructed to encode an analog of a spider silk protein, one of the proteins of the dragline fiber of 35 *Nephila clavipes*. Means are provided whereby such an

artificial gene can be assembled and polymerized to encode a protein of approximately the same length as the natural protein. Further, means are provided whereby such an artificial gene can be expressed in a regulated 5 fashion in a bacterial host, producing large quantities of its protein product. This protein product can be prepared in purified form suitable for forming into a fiber. While the subject of the current invention is a spider silk variant protein, it should be understood 10 that the invention can be extended to encompass other highly repetitive fiber forming proteins or variant forms of such natural proteins.

The present invention provides methods for the production of commercially useful quantities of spider 15 silk proteins in microorganisms, using recombinant DNA technology. Microbial methods of production of such proteins, would provide several advantages. For example microbial sources would provide the basis for production of fiber-forming proteins in large quantities at low 20 enough cost for commercial applications. Microbial hosts would allow the application of recombinant DNA technology for the construction and production of variant forms of fiber-forming proteins, as well as novel proteins that could extend the utility of such 25 fibers. Furthermore, microbial production would permit the rapid preparation of samples of variant proteins for testing. Such proteins would be free of other proteins found in the natural fiber, allowing the properties of the individual proteins to be studied separately.

30 BRIEF DESCRIPTION OF THE DRAWINGS.

SEQUENCE LISTING AND BIOLOGICAL DEPOSITS

Figure 1 illustrates the amino acid sequence (SEQ ID NO.:19) of natural spider dragline protein Spidroin 1 as disclosed by Xu et al., *Proc. Natl. Acad. Sci.* 35 *U.S.A.*, 87, 7120, (1990).

Figure 2 illustrates the amino acid sequences for the monomer (SEQ ID NO.:20) and polymer (SEQ ID NO.:21) of the spider silk DP-1A.9 analog (SEQ ID NO.:80).

Figure 3 illustrates the amino acid sequences for 5 the monomer (SEQ ID NO.:22) and polymer (SEQ ID NO.:23) of the spider silk DP-1B.9 analog (SEQ ID NO.:81).

Figure 4 illustrates the synthetic oligonucleotides L (SEQ ID NOS.:24-26), M1 (SEQ ID NOS.:27-29), M2 (SEQ ID NOS.:30-32), and S (SEQ ID NOS.:33-35) used in the 10 construction of the DNA monomer for DP-1 protein expression.

Figure 5 is a plasmid map illustrating the construction of plasmid pFP510 from pA126i. Plasmid pFP510 is used to construct plasmids for the assembly 15 and polymerization of DNA monomers and genes encoding DP-1A analogs.

Figure 6 is a plasmid map of plasmid pFP202 which is used to construct high level expression vectors.

Figure 7 illustrates the six double stranded 20 synthetic oligonucleotides, A (SEQ ID NOS.:41-43), B (SEQ ID NOS.:44-46), C (SEQ ID NOS.:47-49), D (SEQ ID NOS.:50-52), E (SEQ ID NOS.:53-55), and F (SEQ ID NOS.:56-58), used in the construction of the DNA monomer for DP-2 protein expression.

Figure 8 illustrates the amino acid sequence (SEQ ID NO.:59) of the natural spider silk protein Spidroin 2 as described by Lewis et al. (EP 452925).

Figure 9 illustrates the amino acid sequences of 30 the amino acid monomer (SEQ ID NO.:60) and polymer (SEQ ID NO.:61) of the spider dragline protein 2 analog, DP-2A (SEQ ID NO.:83).

Figure 10 illustrates the amino acid sequences of 35 the amino acid monomer (SEQ ID NO.:62) and polymer (SEQ ID NO.:63) of the spider dragline protein 1 analog, DP-1B.16 (SEQ ID NO.:82).

Figure 11 illustrates the four double stranded synthetic oligonucleotides 1 (SEQ ID NOs.:64-66), 2 (SEQ ID NOs.:67-69), 3 (SEQ ID NOs.:70-72), and 4 (SEQ ID NOs.:73-75) used to construct the synthetic genes 5 encoding DP-1B.16 (SEQ ID NO.:82).

Figure 12 is a plasmid map illustrating the construction of the plasmid pFP206 from pA126i. Plasmid pFP206 was used to construct plasmids used for the assembly and polymerization of the DNA monomer, and 10 genes encoding DP-1B analogs.

Figure 13 illustrates the full nucleic acid sequence (SEQ ID NO.:78) of plasmid pA126i.

Figure 14 illustrates the complete DNA sequence (SEQ ID NO.:79) of pBE346.

15 Figure 15 is a plasmid map illustrating the construction of the plasmid pFP191 which was used to transform *B. subtilis* cells for DP-1A analog protein expression and secretion.

Figure 16 illustrates the four synthetic double-20 stranded oligonucleotides P1, P2, P3, and P4, used to construct the synthetic genes encoding DP-1B.33.

P1 corresponds to SEQ ID NOs.:84, 85, and 86.

P2 corresponds to SEQ ID NOs.:87, 88, and 89.

P3 corresponds to SEQ ID NOs.:90, 91, and 92.

25 P4 corresponds to SEQ ID NOs.:93, 94, and 95.

Figure 17 is a plasmid map of plasmid pHIL-D4, used to construct vectors for intracellular protein expression in *Pichia pastoris*.

Figure 18 is a plasmid map of plasmid pPIC9, used 30 to construct vectors for extracellular protein production in *P. pastoris*.

Figure 19 illustrates the DNA sequence of a portion of plasmid pFO734, an intermediate in the construction of vectors for extracellular protein production in 35 *P. pastoris*.

Figure 20 illustrates DP-1B production by *P. pastoris* strain YFP5028.

Figure 21 illustrates DP-1B production by *P. pastoris* strain YFP5093.

5 Applicants have provided sequence listings 1-107 in conformity with "Rules for the standard representation of nucleotide and amino acid sequence in patent applications" (Annexes I and II to the Decision of the President of the EPO, published in Supplement No. 2 to 10 OJ EPO 12/1992).

Applicants have made the following biological deposits under the terms of the Budapest Treaty.

<u>Deposit or Identification Reference</u>	<u>ATCC Designation</u>	<u>Deposit Date</u>
<i>Escherichia coli</i> , FP 3227	69326	15 June 1993
<i>Escherichia coli</i> , FP 2193	69327	15 June 1993
<i>Escherichia coli</i> , FP 3350	69328	15 June 1993

15 As used herein, the designation "ATCC" refers to the American Type Culture Collection depository located in Rockville, Maryland at 12301 Parklawn Drive, Rockville, MD 20852, U.S.A. The "ATCC No." is the accession number to cultures on deposit at the ATCC.

DETAILED DESCRIPTION OF THE INVENTION

20 The following definitions are used herein and should be referred to for interpretation of the claims and the specification.

25 As used herein, the terms "promoter" and "promoter region" refer to a sequence of DNA, usually upstream of (5' to) the protein coding sequence of a structural gene, which controls the expression of the coding region by providing the recognition for RNA polymerase and/or other factors required for transcription to start at the correct site. Promoter sequences are necessary but not always sufficient to drive the expression of the gene.

A "fragment" constitutes a fraction of the DNA sequence of the particular region.

"Nucleic acid" refers to a molecule which can be single stranded or double stranded, composed of monomers 5 (nucleotides) containing a sugar, phosphate and either a purine or pyrimidine. In bacteria, lower eukaryotes, and in higher animals and plants, "deoxyribonucleic acid" (DNA) refers to the genetic material while "ribonucleic acid" (RNA) is involved in the translation 10 of the information from DNA into proteins.

The terms "peptide", "polypeptide" and "protein" are used interchangeably.

"Regulation" and "regulate" refer to the modulation of gene expression controlled by DNA sequence elements 15 located primarily, but not exclusively upstream of (5' to) the transcription start of a gene. Regulation may result in an all or none response to a stimulation, or it may result in variations in the level of gene expression.

20 The term "coding sequence" refers to that portion of a gene encoding a protein, polypeptide, or a portion thereof, and excluding the regulatory sequences which drive the initiation of transcription. The coding sequence may constitute an uninterrupted coding region 25 or it may include one or more introns bounded by appropriate splice junctions. The coding sequence may be a composite of segments derived from different sources, naturally occurring or synthetic.

The term "construction" or "construct" refers to a 30 plasmid, virus, autonomously replicating sequence, phage or nucleotide sequence, linear or circular, of a single- or double-stranded DNA or RNA, derived from any source, in which a number of nucleotide sequences have been joined or recombined into a unique construction which is 35 capable of introducing a promoter fragment and DNA

sequence for a selected gene product along with appropriate 3' untranslated sequence into a cell.

As used herein, "transformation" is the acquisition of new genes in a cell by the incorporation of nucleic acid.

The term, "operably linked" refers to the chemical fusion of two fragments of DNA in a proper orientation and reading frame to lead to the transcription of functional RNA.

The term "expression" as used herein is intended to mean the transcription and translation to gene product from a gene coding for the sequence of the gene product. In the expression, a DNA chain coding for the sequence of gene product is first transcribed to a complementary RNA which is often a messenger RNA and, then, the thus transcribed messenger RNA is translated into the above-mentioned gene product if the gene product is a protein.

The term "translation initiation signal" refers to a unit of three nucleotides (codon) in a nucleic acid that specifies the initiation of protein synthesis.

The term "signal peptide" refers to an amino terminal polypeptide preceding the secreted mature protein. The signal peptide is cleaved from and is therefore not present in the mature protein. Signal peptides have the function of directing and translocating secreted proteins across cell membranes. The signal peptide is also referred to as signal sequence.

The term "mature protein" refers to the final secreted protein product without any part of the signal peptide attached.

The term "plasmid" or "vector" as used herein refers to an extra-chromosomal element often carrying genes which are not part of the central metabolism of the cell, and usually in the form of circular double-stranded DNA molecules.

The term "restriction endonuclease" refers to an enzyme which catalyzes hydrolytic cleavage within a specific nucleotide sequence in double-stranded DNA.

5 The term "compatible restriction sites" refers to different restriction sites that when cleaved yield nucleotide ends that can be ligated without any additional modification.

10 The term "suitable promoter" will refer to any eukaryotic or prokaryotic promoter capable of driving the expression of a synthetic spider silk variant gene.

The term "spider silk variant protein" will refer to a designed protein, the amino acid sequence of which is based on repetitive sequence motifs and variations thereof that are found in a known a natural spider silk.

15 The term "full length variant protein" will refer to any spider silk variant protein encoded by a synthetic gene which has been constructed by the assembly and polymerization of a DNA monomer.

20 The term "DNA monomer" will refer to a DNA fragment consisting of between 300 and 400 bp which encodes one or more repeating amino acid sequences of a spider silk variant protein. Examples of DNA monomers suitable for the present invention are illustrated in Figures 2, 3, 9 and 10.

25 The term "peptide monomer", "polypeptide monomer" or "amino acid monomer" will refer to the amino acid sequence encoded by a DNA monomer.

30 The term "commercial quantities" will refer to quantities of recombinantly produced desired proteins where at least 1% of the total protein produced by a microbial culture is the desired protein.

The term "desired protein" will refer to any protein considered a valuable product to be obtained from genetically engineered bacteria.

The term "DP-1 analog" will refer to any spider silk variant derived from the amino acid sequence of the natural Protein 1 (Spidroin 1) of *Nephila calvipes* as illustrated in Figure 1.

5 The term "DP-2 analog" will refer to any spider silk variant derived from the amino acid sequence of the natural Protein 2 (Spidroin 2) of *Nephila calvipes* as illustrated in Figure 8.

As used herein the following abbreviations will be
10 used to identify specific amino acids:

<u>Amino Acid</u>	<u>Three-Letter Abbreviation</u>	<u>One-Letter Abbreviation</u>
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Asparagine or aspartic acid	Asx	B
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamine acid	Glu	E
Glutamine or glutamic acid	Glx	Z
Glycine	Gly	G
Histidine	His	H
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

The present invention also provides novel DNA sequences encoding spider silk protein variants that are

suitable for expression of commercial quantities of silk protein in a recombinant host.

It will be appreciated that the advantages of such a protein and such a method are many. Spider silk, especially dragline silk, has a tensile strength of over 200 ksi with an elasticity of nearly 35%, which makes it more difficult to break than either KEVLAR or steel. When spun into fibers, spider silk of the present invention may have application in the bulk clothing industries as well as being applicable for certain kinds of high strength uses such as rope, surgical sutures, flexible tie downs for certain electrical components and even as a biomaterial for implantation (e.g., artificial ligaments or aortic banding). Additionally these fibers may be mixed with various plastics and/or resins to prepare a fiber-reinforced plastic and/or resin product. Furthermore, since spider silk is stable up to 100 °C, these fibers may be used to reinforce thermal injected plastics. These proteins may also be of value in the form of films or coatings. It will be appreciated by one of skill in the art that the properties of the silk fibers may be altered by altering the amino acid sequence of the protein.

The present invention provides a method for the production of analogs of natural spider silk proteins and variants using recombinant DNA technology. The method consists of (1) the design of analog protein sequences based on the amino acid sequence of the fiber forming regions of natural proteins; (2) the design of DNA sequences to encode such analog protein sequences, based on a DNA monomer of at least 50 bp with minimal internal repetitiveness, and making preferential use of codons matched to the preferences of a specific host organism; (3) assembly of the DNA monomer from cloned synthetic oligonucleotides; (4) polymerization of the

DNA monomer to lengths of at least 800 bp, and preferably to lengths approximating the length of the gene encoding the natural protein; (5) inserting the polymerized artificial gene into an appropriate vector 5 able to replicate in the host organism, in such a manner that the gene is operably linked to expression signals whereby its expression can be regulated; (6) producing the protein in the above mentioned microbial host carrying such an expression vector; (7) purifying the 10 protein from the biomass and preparing it in a form suitable for forming into fibers, films, or coatings.

The expression of the desired silk variant protein in *Escherichia coli* is preferred since this host reliably produces high levels of foreign protein and the 15 art is replete with suitable transformation and expression vectors. However, it is not outside the scope of the invention to provide alternative hosts and particularly hosts that facilitate the secretion of the desired protein into the growth medium. Such 20 alternative hosts may include but are not limited to *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Pichia pastoris*, *Aspergillus* spp., *Hansenula* spp., and *Streptomyces* spp. The 25 expression host preferred for the secretion of silk variant protein is *Bacillus subtilis*.

The present invention provides a variety of plasmids or vectors suitable for the cloning of portions of the DNA required for the assembly and expression of the silk variant protein gene in *E. coli*. Suitable 30 vectors for construction contain a selectable marker and sequences allowing autonomous replication or chromosomal integration. Additionally, suitable vectors for expression contain sequences directing transcription and translation of the heterologous DNA fragment. These 35 vectors comprise a region 5' of the heterologous DNA

fragment which harbors transcriptional initiation controls, and optionally a region 3' of the DNA fragment which controls transcriptional termination. It is most preferred when both control regions are derived from 5 genes homologous to *E. coli* although it is to be understood that such control regions need not be derived from the genes native to the specific species chosen as a production host. Suitable vectors can be derived, for example, from a bacteria, a virus (such as bacteriophage 10 T7 or a M-13 derived phage), a cosmid, a yeast or a plant. Protocols for obtaining and using such vectors are known to those in the art. (Sambrook et al., Molecular Cloning: A Laboratory Manual - volumes 1,2,3 (Cold Spring Harbor Laboratory: Cold Spring Harbor, New 15 York, 1989))

Examples of bacteria-derived vectors include plasmid vectors such as pBR322, pUC19, pSP64, pUR278 and pORF1. Illustrative of suitable viral vectors are those derived from phage, vaccinia, retrovirus, baculovirus, 20 or a bovine papilloma virus. Examples of phage vectors include λ^+ , λ EMBL3, 12001, λ gt10, λ gt11, Charon 4a, Charon 40, and λ ZAP/R. pXB3 and pSC11 are exemplary of vaccinia vectors (Chakrabarti et al., *Molec. Cell. Biol.* 5:3401-9 (1985) and Mackett et al., *J. Virol.* 25 49:857864 (1984)). An example of a filamentous phage vector is an M13-derived vector like M13mpl8, and M13mpl9.

For the expression of spider silk variant proteins in *E. coli* bacteria-derived vectors are preferred where 30 plasmids derived from pBR322 are most preferred.

Optionally it may be desired to produce the silk variant protein as a secretion product of a transformed host, such as *B. subtilis*. Secretion of desired proteins into the growth media has the advantage of 35 simplified and less costly purification procedures. It

is well known in the art that secretion signal sequences are often useful in facilitating the active transport of expressible proteins across cell membranes. The creation of a transformed *Bacillus* host capable of 5 secretion may be accomplished by the incorporation of a DNA sequence that codes for a secretion signal functional in the *Bacillus* production host on the expression cassette, between the expression-controlling DNA and the DNA encoding the silk variant protein and in 10 reading frame with the latter. Examples of vectors enabling the secretion of a number of different heterologous proteins by *B. subtilis* have been taught and are described in Nagarajan et al., U.S. Patent 4,801,537; Stephens et al., U.S. Patent 4,769,327; and 15 Biotechnology Handbook 2, *Bacillus*, C. R. Harwood, Ed., Plenum Press, New York (1989).

Secretion vectors of this invention include a regulatable promoter sequence which controls transcription, a sequence for a ribosome binding site 20 which controls translation, and a sequence for a signal peptide which enables translocation of the peptide through the bacterial membrane and the cleavage of the signal peptide from the mature protein. Suitable vectors will be those which are compatible with the 25 bacterium employed. For example, for *B. subtilis* such suitable vectors include *E. coli*-*B. subtilis* shuttle vectors. They will have compatible regulatory sequences and origins of replication. They will be preferably 30 multicopy and have a selective marker gene, for example, a gene coding for antibiotic resistance. An example of such a vector is pTZ18R phagemid, obtainable from Pharmacia, Piscataway, NJ 08854 which confers resistance to ampicillin in *E. coli*. The DNA sequences encoding the promoter, ribosome binding site and signal peptide

may be from any single gene which encodes a secreted product.

The DNA sequences encoding the promoter and ribosome binding site may also be from a different gene 5 than that encoding the signal peptide. The DNA sequences encoding the promoter, ribosome binding site and signal peptide can be isolated by means well known to those in the art and illustrative examples are documented in the literature. See *Biotechnology* 10 *Handbook 2 Bacillus*, C. R. Harwood, Ed., Plenum Press, New York, New York (1989). The promoters in the DNA sequences may be either constitutive or inducible and thus permit the resulting secretion vectors to be differentially regulated.

15 Promoters which are useful to drive expression of heterologous DNA fragments in *E. coli* and *Bacillus* are numerous and familiar to those skilled in the art. Virtually any promoter capable of driving the gene encoding a silk variant protein is suitable for the 20 present invention, where the T7 promoters are preferred in *E. coli* and promoters derived from the *SacB* gene are preferred in *Bacillus*.

Termination control regions may also be derived 25 from various genes native to *E. coli* or *Bacillus* hosts, or optionally other bacterial hosts. It will be appreciated by one of skill in the art that a termination control region may be unnecessary.

For introducing a polynucleotide of the present invention into a bacterial cell, known procedures can be 30 used according to the present invention such as by transformation, e.g., using calcium-permeabilized cells, electroporation, or by transfection using a recombinant phage virus. (Sambrook et al., *Molecular Cloning: A Laboratory Manual* - volumes 1,2,3 (Cold Spring Harbor 35 Laboratory: Cold Spring Harbor, New York, 1989)).

Other known procedures can also be employed to obtain a recombinant host cell that expresses a heterologous spider silk protein according to the present invention, as will be apparent to those skilled in the art.

5 Design of Spider Silk Variant Amino Acid Sequences:

The design of the spider silk variant proteins was based on consensus amino acid sequences derived from the fiber forming regions of the natural spider silk dragline proteins of *Nephila clavipes*. Natural spider 10 dragline consists of two different proteins that are co-spun from the spider's major ampullate gland. The amino acid sequence of both dragline proteins has been disclosed by Xu et al., *Proc. Natl. Acad. Sci. U.S.A.*, 87, 7120, (1990) and Hinman and Lewis, *J. Biol. Chem.*

15 267, 19320 (1992), and will be identified hereinafter as Dragline Protein 1 (DP-1) and Dragline Protein 2 (DP-2).

The amino acid sequence of a fragment of DP-1 is repetitive and rich in glycine and alanine, but is otherwise unlike any previously known amino acid 20 sequence. The repetitive nature of the protein and the pattern of variation among the individual repeats are emphasized by rewriting the sequence as in Figure 1. The "consensus" sequence of a single repeat, viewed in this way, is:

25 A GQG GYG GLG XQG A GRG GLG GQG A GAAAAAAAGG (SEQ ID NO:1)
where X may be S, G, or N.

Examination of Figure 1 shows that individual 30 repeats differ from the consensus according to a pattern which can be generalized as follows: (1) The poly-alanine sequence varies in length from zero to seven residues. (2) When the entire poly-alanine sequence is deleted, so also is the surrounding sequence encompassing AGRGGLGGQQGAGAnGG (SEQ ID NO:2). (3) Aside from the poly-alanine sequence, deletions generally 35 encompass integral multiples of three consecutive

residues. (4) Deletion of GYG is generally accompanied by deletion of GRG in the same repeat. (5) A repeat in which the entire poly-alanine sequence is deleted is generally preceded by a repeat containing six alanine 5 residues.

Synthetic analogs of DP-1 were designed to mimic both the repeating consensus sequence of the natural protein and the pattern of variation among individual repeats. Two analogs of DP-1 were designed and 10 designated DP-1A and DP-1B. DP-1A is composed of a tandemly repeated 101-amino acid sequence listed in Figure 2. The 101-amino acid "monomer" comprises four repeats which differ according to the pattern (1)-(5) above. This 101-amino acid long peptide monomer is 15 repeated from 1 to 16 times in a series of analog proteins. DP-1B was designed by reordering the four repeats within the monomer of DP-1A. This monomer sequence, shown in Figure 3, exhibits all of the regularities of (1)-(5) above. In addition, it exhibits 20 a regularity of the natural sequence which is not shared by DP-1A, namely that a repeat in which both GYG and GRG are deleted is generally preceded by a repeat lacking the entire poly-alanine sequence, with one intervening repeat. The sequence of DP-1B matches the natural 25 sequence more closely over a more extended segment than does DP-1A.

The amino acid sequence of a fragment of DP-2 is also repetitive and also rich in glycine and alanine, but is otherwise unlike any previously known amino acid 30 sequence, and, aside from a region of consecutive alanine residues, different from DP-1. The repetitive nature of the protein and the pattern of variation among the individual repeats are emphasized by rewriting the sequence as in Figure 8. The "consensus" sequence of a 35 single repeat, viewed in this way, is:

[GPGGY GPGQQ]₃ GPSGPGS A₁₀ (SEQ ID NO:18)

Examination of Figure 8 shows that individual repeats differ from the consensus according to a pattern which can be generalized as follows: (1) The polyalanine-rich sequence varies in length from six to ten residues. (2) Aside from the poly-alanine sequence, individual repeats differ from the consensus repeat sequence by deletions of integral multiples of five consecutive residues consisting of one or both of the pentapeptide sequences GPGGY (SEQ ID NO:3) or GPGQQ (SEQ ID NO:4).

Synthetic analogs of DP-2 were designed to mimic both the repeating consensus sequence of the natural protein and the pattern of variation among individual repeats. The analog DP-2A is composed of a tandemly repeated 119-amino acid sequence listed in Figure 9. The 119-amino acid "peptide monomer" comprises three repeats which differ according to the pattern (1)-(2) above. This 119-amino acid long peptide monomer is repeated from 1 to 16 times in a series of analog proteins.

Design of DNA encoding Spider Silk Variant Proteins:

DNA sequences encoding the designed analog amino acid sequences were devised according to the following criteria: (1) The DNA monomer was to be at least 300 bp in length; (2) within the monomer, repetitiveness of the sequence was minimized, with no repeated sequence longer than 17 bp and minimal repetitiveness of sequences longer than 10 bp; (3) where possible, codons were chosen from among the codons found preferentially in highly expressed genes of the intended host organism (*E. coli*) with preference for codons providing balanced A+T/G+C base ratios; and (4) predicted secondary structure of mRNA within the monomer was dominated by long-range interactions rather than shorter range base

pairing. No attempt was made to minimize secondary structure of the mRNA.

Assembly of DP-1 and DP-2 Analog Genes:

Assembly of the synthetic dragline analog genes was 5 accomplished by first assembling the appropriate DNA monomers followed by polymerization of these monomers to form the completed gene.

Synthetic DNA monomers, based on the consensus peptide monomers described above were assembled from 10 four to six cloned double stranded synthetic oligonucleotides. Each oligonucleotide was designed to encode a different portion of the the peptide monomer. Briefly, the oligonucleotides were each cloned into separate suitable plasmid vectors containing an 15 ampicillin resistance gene. A suitable *E. coli* host was transformed with the plasmids and screened for the presence of the correct vector by standard methods. After the oligonucleotides were cloned the DNA monomer was sequentially assembled. Vectors containing 20 individual oligonucleotides were digested and the plasmid DNA was purified by gel electrophoresis. Purified plasmid DNA containing two different 25 oligonucleotide sequences were then incubated under ligating conditions and the ligation products were used to transform a suitable *E. coli* host. These transformants comprised two of the oligonucleotide sequences linked in tandem. A similar procedure was followed for the creation of the full DNA monomer, comprising four to six of the oligonucleotides. 30 Additional confirmation of the existence of the correct DNA insertions was obtained by direct DNA sequencing. The present invention provides several DNA monomers useful for the production of DP-1A and DP-1B analogs. In general DNA monomers used to produce the the analog

DP-1B.16 are preferred since this construct avoids codons rarely used by the *E. coli* production host.

The assembled DNA monomer was then polymerized by a method essentially as described by Kempe et al. (Gene 5 39, 239, (1985)). This method consists of a series of successive doublings of the sequence of interest.

Briefly, the DNA monomer containing the cloned oligonucleotides was digested with suitable restriction enzymes and incubated under annealing conditions

10 followed by ligation to produce a series of constructs containing multiple repeats of the monomer. Ligation products were used to transform a suitable *E. coli* host and intact plasmids were selected on the basis of ampicillin resistance. Subsequent analysis of plasmid 15 DNA by gel electrophoresis resulted in the identification of transformants containing plasmids with 2, 4, 8, and 16 tandem repeats of the DNA monomer.

These protein products were analyzed by SDS polyacrylamide gel electrophoresis and detected and 20 quantitated by immunochemical staining using a polyclonal antiserum raised in rabbits against a synthetic peptide analogous to a fragment of the natural protein.

Expression and purification of Protein:

25 High level expression of the spider dragline protein analogs in *E. coli* was achieved by inserting the synthetic genes into plasmid vectors pFP202 and pFP204, which were derived from the well-known vector pET11a. In these vectors, the dragline protein-coding gene is 30 inserted in such a manner as to be operably linked to a promoter derived from bacteriophage T7. This promoter is joined with sequences derived from the lac operator of *E. coli*, which confers regulation by lactose or analogs (IPTG). The *E. coli* host strain BL21(DE3) 35 contains a lambda prophage which carries a gene encoding

bacteriophage T7 RNA polymerase. This gene is controlled by a promoter which is also regulated by lactose or analogs. In addition to the phage T7 promoter, the vectors pFP202 and pFP204 provide 5 sequences which encode a C-terminal tail containing six consecutive histidine residues appended to the dragline protein-coding sequences. This tail provides a means of affinity purification of the protein under denaturing conditions through its adsorption to resins bearing 10 immobilized Ni ions.

DP-1 analog protein was produced by *E. coli* at levels of approximately 5-20% of total protein. Of this, approximately 20-40% was recovered in purified form as full-length protein. DP-2 analog protein was 15 produced at approximately 5% of total cell protein, of which approximately 30% was recovered in purified form as full-length protein.

The following examples are meant to illustrate the invention but should not be construed as limiting it in 20 any way.

EXAMPLES

GENERAL METHODS

The position of the newly engineered restriction sites is indicated in the figures and any one skilled in 25 the art can repeat these constructs with the available information.

The source of the genes and the various vectors described throughout this application are as follows.

The anti-DP-1 and anti-DP-2 antisera were prepared 30 by Multiple Peptide Systems, San Diego, CA.

Restriction enzyme digestions, phosphorylations, ligation, transformations and other suitable methods of genetic engineering employed herein are described in Sambrook et al., *Molecular Cloning: A Laboratory* 35 Manual - volumes 1,2,3 (Cold Spring Harbor Laboratory:

Cold Spring Harbor, New York, 1989), and in the instructions accompanying commercially available kits for genetic engineering.

Bacterial cultures and plasmids to carry out the 5 present invention are available either commercially (from Novagen, Inc., Madison, WI) or from the *E. coli* Genetic Stock Center, Yale University, New Haven, CT, the *Bacillus* Genetic Stock Center, Ohio State University, Columbus, OH, or the ATCC and, along with 10 their sources, are identified in the text and examples which follow. Unless otherwise specified standard reagents and solutions used in the following examples were supplied by Sigma Chemical Co. (St. Louis, MO)

Isolation of restriction fragments from agarose 15 gels used the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA), and was performed as specified by the manufacturer.

EXAMPLE 1

CONSTRUCTION OF THE SYNTHETIC GENES

20 DP-1A.9 AND DP-1B.9

Oligonucleotide design and cloning:

Synthetic genes encoding DP-1A.9 and DP-1B.9 were assembled from four double stranded synthetic oligonucleotides labeled L (SEQ ID NOS.:24, 25, and 26), 25 M1 (SEQ ID NOS.:27, 28, and 29), M2 (SEQ ID NOS.:30, 31, and 3), and S (SEQ ID NOS.:33, 34, and 35) whose sequences are shown in Figure 4. The oligonucleotides were provided by the manufacturer (Midland Certified Reagents, Midland, TX) in double stranded form with 30 5'-OH groups phosphorylated. Methods of oligonucleotide synthesis, purification, phosphorylation, and annealing to the double stranded form are well known to those skilled in the art.

The four double stranded oligonucleotides were 35 separately cloned by inserting them into a plasmid

vector pFP510 (Figure 5). This vector was derived from the plasmid pA126i (see Figure 13), the complete nucleotide sequence of which is provided in SEQ ID NO.:78 and Figure 13. Details of the structure of pA126i are not important for the construction, aside from the following essential features: (a) a replication origin active in *E. coli*; (b) a selectable genetic marker, in this case a gene conferring resistance to the antibiotic ampicillin; (c) sites for restriction endonucleases BamHI and BglII with no essential sequences between them; and (d) a third restriction site (PstI), located within the selectable marker, which produces cohesive ends incompatible with those produced by BamHI and BglII. For the construction of pFP510, DNA of plasmid pA126i was digested with endonucleases BamHI and BglII, then recovered by adsorption to glass beads in the presence of NaI GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). To approximately 0.1 pmole of the eluted plasmid DNA was added 10 pmoles of the double stranded, phosphorylated oligonucleotide SF4/5 (Figure 5). The mixture was incubated under ligation conditions with T4 polynucleotide ligase for 19 h at 4 °C. Ligated DNA was then digested with endonuclease XmaI to linearize any remaining parental pA126i and used to transform *E. coli* SK2267 (obtained from the *E. coli* Genetic Stock Center, Yale University, New Haven, CT) which had been made competent by calcium treatment as described by Sambrook et al., op. cit. Plasmid DNA isolated from ampicillin resistant transformants was characterized by digestion separately with endonucleases ApaI and BamHI, and a transformant containing the desired plasmid was identified and designated pFP510.

DNA of plasmid pFP510 was digested with endo-
35 nucleases SfiI and DraIII and purified by the GENECLEAN®

procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). To approximately 0.1 pmole of the eluted plasmid DNA was added 10 pmoles of one of the double stranded, phosphorylated oligonucleotides L, M1, M2, or S (Figure 4). The four plasmid-oligonucleotide mixtures were incubated under ligation conditions for 15 h at 4 °C, then for 20 min at 23 °C and finally ligation was terminated by incubation for 3 min at 65 °C. Aliquots of ligated DNA were used to transform *E. coli* SK2267 and ampicillin resistant transformants were selected.

Clones containing oligonucleotides L, M1, and M2 shown in Figure 4 were identified by screening plasmid DNA isolated from individual transformants with endonuclease AlwNI, a recognition site for which is present in the oligonucleotides. Clones containing oligonucleotide S were identified by screening plasmid DNA isolated from individual transformants with endonucleases BglI and DraIII. Plasmid DNA from putative clones was further characterized by digestion with endonucleases EcoRI, SfiI, and DraIII in order to establish that the oligonucleotide sequences were oriented correctly in the plasmid. The inserts were excised with endonucleases BamHI and BglII and analyzed by electrophoresis in 4% NuSieve agarose (FMC) to verify that the plasmid had acquired only a single copy of the oligonucleotide. Correct clones were identified and their plasmids were designated pFP521 (oligonucleotide L), pFP533 (oligonucleotide M1), pFP523 (oligonucleotide M2), and pFP524 (oligonucleotide S). DNA sequences of all four cloned oligonucleotides were verified by DNA sequencing.

DNA sequencing was carried out essentially according to procedures provided by the supplier (U.S. Biochemicals) with the Sequenase 2.0 kit for DNA sequencing with 7-deaza-GTP. Plasmid DNA was prepared using the Magic Minipreps kit (Promega). Template DNA

was denatured by incubating 20 μ l miniprep DNA in 40 μ l (total volume) 0.2 M NaOH for 5 min at 23 °C. The mixture was neutralized by adding 6 μ l 2 M ammonium acetate (adjusted to pH 4.5 with acetic acid), and the 5 DNA was precipitated by adding 0.15 mL ethanol, recovered by centrifugation, washed with cold 70% ethanol, and vacuum dried. Primers for sequencing were as follows:

10 SI1: 5'-ACGACCTCATCTAT (SEQ ID NO:5)
 SI5: 5'-CTGCCTCTGTCATC (SEQ ID NO:6)
 SI20: 5'-AATAGGCGTATCAC (SEQ ID NO:7)

Primers SI1 and SI5 anneal to sites on opposite strands in pA126i. SI5 primes synthesis into the sequences of interest from 31 bp beyond the BamHI site. SI1 primes 15 synthesis on the opposite strand into the sequences of interest from 38 bp beyond the BglII site. For sequencing in the vector pFP206 (see below) the primer SI20, which anneals 25 bp beyond the BglII site, was substituted for SI1 (Figure 12). Polyacrylamide gels 20 for DNA sequencing were run at 52 °C.

Assembly of the Gene:

For assembly of subsequence M2L, plasmid pFP523 (M2) was digested with endonucleases PstI and DraIII, and plasmid pFP521 (L) was digested with endonucleases 25 PstI and SfiI. Digested plasmid DNA was fractionated by electrophoresis in a 1.2% agarose (low melting, BioRad) gel. Ethidium bromide-stained bands containing the oligonucleotide sequences, identified by their relative sizes, were excised, the excised bands combined, and the 30 DNA recovered from melted agarose by the GENECLEAN® procedure (Biol01, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined DNA fragments were incubated under ligation conditions and an aliquot was used to transform *E. coli* W3110 (available from the *E. coli* Genetic Stock 35 Center, Yale University, New Haven, CT.). Ampicillin

resistant transformants were selected. Plasmid DNA was isolated from several transformants, digested with endonucleases BamHI and BglII, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the 5 expected size was identified and designated pFP525.

Assembly of subsequence M1S was accomplished in the same manner, starting with plasmids pFP533 (digested with PstI and DraIII) and pFP524 (digested with PstI and SfiI). Plasmid containing the M1S subsequence was 10 identified and designated pFP531.

For assembly of the DNA monomer (M2LM1S), plasmid pFP525 (M2L) was digested with endonucleases PstI and DraIII, and plasmid pFP531 (M1S) was digested with endonucleases PstI and SfiI. Digested plasmid DNA was 15 fractionated by electrophoresis in a 1.2% low melting agarose gel. Ethidium bromide-stained bands containing the M2L and M1S sequences, respectively, identified by their relative sizes, were excised, the excised bands combined, and the DNA recovered from melted agarose by 20 the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined DNA fragments were incubated under ligation conditions and an aliquot was used to transform *E. coli* W3110. Ampicillin resistant transformants were selected. Plasmid DNA was isolated 25 from several transformants, digested with endonucleases BamHI and BglII, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was identified and designated pFP534. The DNA inserts in plasmids pFP523, pFP521, pFP533, pFP524, 30 pFP525, pFP531, and pFP534 were verified by direct DNA sequencing as previously described.

Polymerization of the Gene:

The synthetic gene was extended by sequential doubling, starting with the monomer sequence in pFP534. 35 For doubling any insert sequence, an aliquot of plasmid

DNA was digested with endonucleases *Pst*I and *Dra*III, and a separate aliquot of the same plasmid was digested with endonucleases *Pst*I and *Sfi*I. Digests were fractionated by electrophoresis on low melting agarose, and ethidium bromide stained fragments containing insert sequences were identified by their relative sizes. In some cases, the two fragments were not adequately separated, so it was necessary to cut the non-insert-containing fragment with a third enzyme, usually *Mlu*I.

10 Each of the two insert sequence-containing fragments has one end generated by endonuclease *Pst*I. Annealing of these compatible single stranded ends and ligation results in reconstitution of the gene that confers ampicillin resistance, part of which is carried 15 on each fragment. The other end of each fragment displays a single stranded sequence generated by either *Dra*III or *Sfi*I. These sequences are, by design, complementary, and annealing and ligation results in a head-to-tail coupling of two insert sequences, with 20 concomitant loss of both sites at the junction. The principle of this method of insert sequence doubling was described by Kempe et al. (Gene 39, 239-245 (1985)).

The two insert-containing fragments, purified by electrophoresis and recovered by the GENECLEAN® 25 procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA), were combined and incubated under ligation conditions. An aliquot was used to transform *E. coli* W3110. Ampicillin resistant transformants were selected. Plasmid DNA was isolated from several transformants, 30 digested with endonucleases *Bam*HI and *Bgl*II, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was identified.

By this procedure a series of plasmids was constructed containing 2, 4, 8, and 16 tandem repeats of 35 the DNA monomer sequence M2LM1S, encoding the series of

DP-1A analogs. In addition, analogous methods were used to construct genes encoding the series of DP-1B analogs. For this purpose, subsequences SL (from pFP524 and pFP521) and M1M2 (from pFP533 and pFP523) were first 5 constructed, then combined to form the monomer SLM1M2, which was polymerized as described. It should be apparent that similar methods can be used to assemble any combination of subsequences carried in the vector pFP510, or any other appropriate vector, provided that 10 the subsequences are bounded by cleavage sites for restriction endonucleases that generate compatible ends (complementary single stranded ends or blunt ends). In addition to various monomer sequences, polymers of any number of repeats of the monomer sequence can be 15 assembled in the same way, starting with plasmids containing inserts of different sizes.

EXAMPLE 2

SYNTHETIC GENE DP-1B.16

A second set of genes encoding DP-1B, designated 20 DP-1B.16 (SEQ ID NO.:82), were designed to reduce the number of codons which are rarely used in highly expressed *E. coli* genes, but at the same time encoding proteins of the same repeating sequence. The sequence of the DP-1B.16 peptide monomer is shown in Fig. 10 and 25 in SEQ ID NO.:82.

Oligonucleotide Synthesis and Cloning:

Synthetic genes encoding DP-1B.16 (SEQ ID NO.:82) were assembled from four double stranded synthetic oligonucleotides whose sequences (SEQ ID NOs.:64, 65, 30 66; SEQ ID NOs.:67, 68, 69; SEQ ID NOs.:70, 71, 72; and SEQ ID NOs.:73, 74, 75) are shown in Figure 11. The oligonucleotides were provided by the manufacturer (Midland Certified Reagents, Midland, TX) in single stranded form with 5'-OH groups not phosphorylated. For 35 annealing to the double stranded form, complementary

single stranded oligonucleotides (667 pmoles each) were mixed in 0.2 mL buffer containing 0.01 M Tris-HCl, 0.01 M MgCl₂, 0.05 M NaCl, 0.001 M dithiothreitol, pH 7.9. The mixture was heated in boiling water for 1 minute, then allowed to cool slowly to 23 °C over approximately 3 h.

The four double stranded oligonucleotides were separately cloned by inserting them into a plasmid vector pFP206 (Figure 12). This vector was derived from 10 the plasmid pA126i as illustrated in Fig. 12. Briefly, DNA of plasmid pA126i was digested with endonucleases BamHI and EcoRI, and the two fragments were separated by electrophoresis in a 1.2% agarose (low melting, BioRad). The larger of the two fragments was excised from the 15 ethidium bromide-stained gel and recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). To approximately 0.1 pmole of the eluted DNA fragment was added 10 pmoles of the double stranded, phosphorylated oligonucleotide SF31/32 (Figure 12). The 20 mixture was incubated under ligation conditions with T4 polynucleotide ligase for 8.5 h at 4 °C. Ligated DNA was used to transform *E. coli* HB101, which had been made competent by calcium treatment. Plasmid DNA isolated from ampicillin resistant transformants was 25 characterized by digestion separately with endonucleases HindIII, EcoRI, BglII, and BamHI, and a transformant containing the desired plasmid was identified and designated pFP206.

DNA of plasmid pFP206 was digested with 30 endonucleases BamHI and BglII and purified by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). To approximately 0.1 pmole of the eluted plasmid DNA was added 10 pmoles of one of the double stranded oligonucleotides 1 (SEQ ID NOS.:64, 65, 66) 2 35 (SEQ ID NOS.:67, 68, 69), 3 (SEQ ID NOS.:70, 71, 72), or

4 (SEQ ID NOS.:73, 74, 75). The four plasmid-oligonucleotide mixtures were incubated under ligation conditions for 15 h at 4 °C, then ligation was terminated by incubation for 3 min at 70 °C. Ligated 5 DNA was then digested with endonuclease HindIII to linearize any remaining parental pFP206. Aliquots of ligated DNA were used to transform *E. coli* HB101 and ampicillin resistant transformants were selected. Clones containing oligonucleotides 1, 2, 3, or 4 were 10 identified by screening plasmid DNA isolated from individual transformants with endonucleases BamHI and PstI. In plasmids with inserts in the desired orientation, the shorter of two BamHI-PstI fragments of pFP206 is lengthened by the length of the cloned 15 oligonucleotide. Plasmid DNA from putative clones was further characterized by digestion with endonucleases BamHI and BglII and analysis by electrophoresis in 3% NuSieve agarose (FMC), 1% Agarose (Sigma Chemical Co.) to verify that the plasmid had acquired only a single 20 copy of the oligonucleotide in the correct orientation. Correct clones were identified and their plasmids were designated pFP636 (oligonucleotide 1), pFP620 (oligonucleotide 2), pFP641 (oligonucleotide 3), and pFP631 (oligonucleotide 4). Sequences of all four 25 cloned oligonucleotides were verified by DNA sequencing as described above.

Assembly of the Gene:

For assembly of subsequence 1,2, plasmid pFP636 (1) was digested with endonucleases PstI and BamHI, and 30 plasmid pFP620 (2) was digested with endonucleases PstI and BglII. Digested plasmid DNA was fractionated by electrophoresis in a 1.2% agarose (low melting, BioRad) gel. Ethidium bromide-stained bands containing the oligonucleotide sequences, identified by their relative 35 sizes, were excised, the excised bands combined, and the

DNA recovered from melted agarose by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined DNA fragments were incubated under ligation conditions and an aliquot was used to transform 5 *E. coli* HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated from several transformants, digested with endonucleases BamHI and BglIII, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was 10 identified and designated pFP647.

Assembly of subsequence 3,4 was accomplished in the same manner, starting with plasmids pFP641 (digested with PstI and BamHI) and pFP631 (digested with PstI and BglIII). Plasmid containing the 3,4 subsequence was 15 identified and designated pFP649.

For assembly of the DNA monomer (1,2,3,4), plasmid pFP647 (1,2) was digested with endonucleases PstI and BamHI, and plasmid pFP640 (3,4) was digested with endonucleases PstI and BglIII. Digested plasmid DNA was 20 fractionated by electrophoresis in a 1.2% low melting agarose gel. Ethidium bromide-stained bands containing the 1,2 and 3,4 sequences, respectively, identified by their relative sizes, were excised, the excised bands combined, and the DNA recovered from melted agarose by 25 the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined DNA fragments were incubated under ligation conditions and an aliquot was used to transform *E. coli* HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated 30 from several transformants, digested with endonucleases BamHI and BglIII, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was identified and designated pFP652. The DNA insert in plasmid pFP652 was verified by direct DNA 35 sequencing as described above.

Polymerization of the Gene:

The synthetic gene was extended by sequential doubling, starting with the monomer sequence in pFP652. For doubling any insert sequence, an aliquot of plasmid 5 DNA was digested with endonucleases PstI and BamHI, and a separate aliquot of the same plasmid was digested with endonucleases PstI and BglII. Digests were fractionated by electrophoresis on low melting agarose, and ethidium bromide stained fragments containing insert sequences 10 were identified by their relative sizes. The two insert-containing fragments, purified by electrophoresis and recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA), were combined and 15 incubated under ligation conditions. At the third doubling, the two fragments in the BamHI digest were not adequately separated, so the eluted band contained both fragments. In this case a two-fold excess of the BglII-PstI fragment was used in the ligation. An aliquot of the ligated DNA was used to transform *E. coli* 20 HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated from several transformants, digested with endonucleases BamHI and BglII, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was 25 identified.

By this procedure a series of plasmids was constructed containing 2, 4, 8, and 16 tandem repeats of the DNA monomer sequence 1 (SEQ ID NOS.:64, 65, 66), 2 (SEQ ID NOS.:67, 68, 69), 3 (SEQ ID NOS.:70, 71, 72), 30 4 (SEQ ID NOS.:73, 74, 75), encoding the series of DP-1B.16 analogs. These plasmids were designated pFP656 (2 repeats), pFP661 (4 repeats), pFP662 (8 repeats), and pFP665 (16 repeats), respectively.

EXAMPLE 3SYNTHETIC GENE DP-2AOligonucleotide Synthesis and Cloning:

Synthetic genes encoding DP-2A were assembled from 5 six double stranded synthetic oligonucleotides whose sequences are shown in Figure 7. The oligonucleotides were provided by the manufacturer (Midland Certified Reagents, Midland, TX) in double stranded form with 10 5'-OH groups not phosphorylated. The six double stranded oligonucleotides were separately cloned by inserting them into the plasmid vector pFP206.

DNA of plasmid pFP206 was digested with 15 endonucleases BamHI and BglII and purified by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). To approximately 0.1 pmole of the eluted plasmid DNA was added 10 pmoles of one of the double stranded oligonucleotides A (SEQ ID NOS.:41, 42, 43), B (SEQ ID NOS.:44, 45, 46), C (SEQ ID NOS.:47, 48, 49), D (SEQ ID NOS.:50, 51, 52), E (SEQ ID NOS.:53, 54, 55), 20 or F (SEQ ID NOS.:56, 57, 58). The six plasmid-oligonucleotide mixtures were incubated under ligation conditions for 15 h at 4 °C, then ligation was terminated by incubation for 3 min at 70 °C. Ligated DNA was then digested with endonuclease HindIII to 25 linearize any remaining parental pFP206. Aliquots of ligated DNA were used to transform *E. coli* HB101 and ampicillin resistant transformants were selected. Clones containing oligonucleotides A, B, C, D, E, or F were identified by screening plasmid DNA isolated from 30 individual transformants with endonucleases BamHI and PstI. In plasmids with inserts in the desired orientation, the shorter of two BamHI-PstI fragments of pFP206 is lengthened by the length of the cloned oligonucleotide. Plasmid DNA from putative clones was 35 further characterized by digestion with endonucleases

BamHI and BglIII and analysis by electrophoresis in 3% NUSIEVE agarose (FMC), 1% Agarose (Sigma Chemical Co.) to verify that the plasmid had acquired only a single copy of the oligonucleotide in the correct orientation.

5 Correct clones were identified and their plasmids were designated pFP193 (oligonucleotide A), pFP194 (oligonucleotide B), pFP195 (oligonucleotide C), pFP196 (oligonucleotide D), pFP197 (oligonucleotide E), and pFP198 (oligonucleotide F).

10 Assembly of the Gene:

For assembly of subsequence AB, plasmid pFP193 (A) was digested with endonucleases PstI and PvuII, and plasmid pFP194 (B) was digested with endonucleases PstI and SmaI. Digested plasmid DNA was fractionated by 15 electrophoresis in a 1.2% agarose (low melting, BioRad) gel. Ethidium bromide-stained bands containing the oligonucleotide sequences, identified by their relative sizes, were excised, the excised bands combined, and the DNA recovered from melted agarose by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined DNA fragments were incubated under ligation conditions and an aliquot was used to transform 20 *E. coli* HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated from several 25 transformants, digested with endonucleases BamHI and BglIII, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was identified and designated pFP300 (AB).

Assembly of subsequence CD was accomplished in the 30 same manner, starting with plasmids pFP195 (digested with PstI and SnaBI) and pFP196 (digested with PstI and SmaI). Plasmid containing the CD subsequence was identified and designated pFP578. Assembly of subsequence EF was accomplished in the same manner, 35 starting with plasmids pFP197 (digested with PstI and

SnaBI) and pFP198 (digested with PstI and SmaI). Plasmid containing the EF subsequence was identified and designated pFP583. The DNA inserts in plasmids pFP300, pFP578, and pFP583 were verified by direct DNA sequencing as described above.

Assembly of subsequence CDEF was accomplished similarly, starting with plasmids pFP578 (digested with PstI and PvuII) and pFP583 (digested with PstI and SmaI). Plasmid containing the CDEF subsequence was identified and designated pFP588.

For assembly of the DNA monomer (ABCDEF), plasmid pFP300 (AB) was digested with endonucleases PstI and PvuII, and plasmid pFP588 (CDEF) was digested with endonucleases PstI and SmaI. Digested plasmid DNA was fractionated by electrophoresis in a 1.2% low melting agarose gel. Ethidium bromide-stained bands containing the AB and CDEF sequences, respectively, identified by their relative sizes, were excised, the excised bands combined, and the DNA recovered from melted agarose by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined DNA fragments were incubated under ligation conditions and an aliquot was used to transform *E. coli* HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated from several transformants, digested with endonucleases BamHI and BglII, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was identified and designated pFP303. The DNA insert in plasmid pFP303 was verified by direct DNA sequencing.

Polymerization of the Gene:

The synthetic gene was extended by sequential doubling, starting with the monomer sequence in pFP303. For doubling any insert sequence, an aliquot of plasmid DNA was digested with endonucleases PstI and PvuII, and

a separate aliquot of the same plasmid was digested with endonucleases *Pst*I and *Sma*I. Digests were fractionated by electrophoresis on low melting agarose, and ethidium bromide stained fragments containing insert sequences 5 were identified by their relative sizes. The two insert-containing fragments, purified by electrophoresis and recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA), were combined and 10 incubated under ligation conditions. An aliquot of the ligated DNA was used to transform *E. coli* HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated from several transformants, 15 digested with endonucleases *Bam*HI and *Bgl*III, and analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was identified.

By this procedure a series of plasmids was constructed containing 2, 4, 8, and 16 tandem repeats of the DNA monomer sequence ABCDEF, encoding the series of DP-2A analogs. These plasmids were designated pFP304 (2 20 repeats), pFP596 (4 repeats), pFP597 (8 repeats), and pFP598 (16 repeats), respectively.

EXAMPLE 4

EXPRESSION OF DP-1 AND DP-2 ANALOG GENES IN E. COLI
Immunoassay

25 For detection of DP-1 analog amino acid sequences, polyclonal antisera were raised in rabbits by immunization with a synthetic peptide matching the most highly conserved segment of the consensus repeat sequence of the natural protein. The peptide (sequence 30 CGAGQQGGYGGGLGSQGAGRG-NH₂) (SEQ ID NO:8) was synthesized by standard solid phase methods (Multiple Peptide Systems, San Diego, CA) and coupled through its terminal Cys thiol to Keyhole Limpet Hemocyanin via maleimido-benzoyl-N-hydroxysuccinimide ester. Similarly, for 35 detection of DP-2 analog amino acid sequences, antisera

were raised against a peptide of sequence CGPGQQGPGGGYGPGQQGPS-NH₂ (SEQ ID NO:9), which reflects the consensus repeat sequence of the natural protein DP-2.

5 For the growth of cultures to assess production levels, 20 mL L broth (per liter: 10 g Bacto-Tryptone (Difco), 5 g Bacto-Yeast Extract (Difco), 5 g NaCl, pH adjusted to 7.0 with NaOH) containing 0.1 mg/mL ampicillin in a 125 mL baffled Erlenmeyer flask was
10 inoculated at an absorption (A₆₀₀ nm) of approximately 0.05 with cells eluted from an L-agar plate containing 0.1 mg/mL ampicillin, which had been grown overnight at 37 °C. The culture was shaken at 37 °C until the A₆₀₀ nm reached approximately 1.0, at which time IPTG was added
15 to a final concentration of 1 mM. Samples (0.5 mL) were taken immediately before IPTG addition and after an additional 3 h at 37 °C. Cells were immediately recovered by centrifugation in a microfuge, supernatant was removed, and the cell pellet was frozen in dry ice
20 and stored at -70 °C.

For analysis by polyacrylamide gel electrophoresis, cell pellets were thawed, suspended in 0.2 mL sample preparation buffer (0.0625 M Tris-HCl, pH 6.8, 2% w/v Na-dodecyl sulfate, 0.0025% w/v bromphenol blue, 10% v/v glycerol, 2.5% v/v 2-mercaptoethanol), and incubated in a boiling water bath for 5 min. Aliquots (15 µl) were applied to a 4-12% gradient polyacrylamide gel (Novex) and subjected to electrophoresis until the dye front was less than 1-cm from the bottom of the gel. The gel was
25 stained with Coomassie Brilliant Blue. A second gel (6% acrylamide) was run with similar samples, then protein bands were transferred electrophoretically to a sheet of nitrocellulose, using an apparatus manufactured by Idea
30 Scientific, Inc. The buffer for transfer contained (per

liter) 3.03-g Trishydroxymethyl aminomethane, 14.4-g glycine, 0.1% w/v SDS, 25% v/v methanol.

The nitrocellulose blot was stained immuno-chemically as follows. Protein binding sites on the 5 sheet were blocked by incubation with "Blotto" (3% nonfat dry milk, 0.05% TWEEN 20, in Tris-saline (0.1 M Tris-HCl, pH 8.0, 0.9% w/v NaCl)) for 30 min at room temperature on a rocking platform. The blot was then 10 incubated for 1 h with anti DP-1 serum or anti DP-2 serum, diluted 1:1000 in "Blotto", washed with Tris-saline, and incubated for 1 h with horseradish peroxidase-conjugated goat anti-rabbit IgG serum (Kierkegaard and Perry Laboratories, Gaithersburg, MD), diluted 1:1000 in "Blotto". After again washing with 15 Tris-saline, the blot was exposed to a solution of 18 mg 4-chloro-1-naphthol in 6 mL methanol, to which had been added 24 mL Tris-saline and 30 μ l 30% H₂O₂.

For quantitation of DP-1 antigen production, cell extracts were prepared by either of two procedures.

20 Procedure 1: The cell pellet from 0.5 mL culture was resuspended in 0.084 mL 50 mM EDTA, pH 8.0, to which was then added 10 μ l 10 mg/mL egg white lysozyme in the same buffer, 1 μ l 2 mg/mL bovine pancreatic ribonuclease, and 5 μ l 0.1 M phenyl methane sulfonyl 25 fluoride in ethanol. After 15 min at 37 °C, 1 μ l 1 mg/mL DNase I was added, along with 3 μ l 1 M MgCl₂, 1 M MgSO₄, and incubation was continued for 10 min at 37 °C. The resulting lysate was clarified by centrifugation for 5 min in a microfuge, and the 30 supernatant was diluted to 0.5 mL with Tris-saline.

Procedure 2: The cell pellet was resuspended in 0.5 mL of buffer 8.0G containing 6 M guanidine-HCl, 0.1 M NaH₂PO₄, 0.01 M Tris-HCl, 5 mM 2-mercaptoethanol, pH adjusted to 8.0 with NaOH. After thorough mixing and

incubation for 1 h at 23 °C, cell debris was removed by centrifugation for 15 minutes in a microfuge.

Aliquots (1 μ l) of serial dilutions in Tris saline (Procedure 1) or buffer 8.0G (Procedure 2) were spotted 5 onto nitrocellulose, along with various concentrations of a standard solution of purified DP-1 8-mer (8 repeats of 101 amino acid residues). The nitrocellulose sheet was then treated as described above for the Western blot. The concentration of DP-1 antigen in each sample 10 was estimated by matching the color intensity of one of the standard spots.

Production strains:

Vectors:

To construct bacterial strains for production of 15 DP-1, cloned synthetic DP-1-coding DNA sequences were inserted into plasmid vector, pFP202 (Figure 6) or pFP204, which were derived from plasmid pFP200, which was, in turn, derived from the plasmids pET11a and pET9a of Studier et al., *Methods in Enzymology*, 185, 60 20 (1990). Plasmids pET9a and pET11a and host strains BL21, BL21(DE3), HMS174, and HMS174(DE3) were obtained from Novagen, Madison, WI.

To construct the plasmid pFP200, DNA of plasmids pET9a and pET11a were digested with endonucleases EcoRI 25 and AlwNI and the digests fractionated separately by electrophoresis in low-melting agarose. The appropriate ethidium bromide-stained bands (from pET9a, the band carrying the gene that confers resistance to kanamycin, and from pET11a, the band carrying the T7 promoter) were 30 identified by size, excised and recovered from melted gel slices by the GENECLEAN® procedure (Biol01, Inc., P.O. Box 2284, La Jolla, CA). Equivalent amounts of the purified DNA bands were combined and incubated under ligation conditions. An aliquot of the ligated DNA was 35 used to transform *E. coli* BL21 and transformants were

selected for resistance to kanamycin (50 µg/mL).

Plasmid DNA from individual transformants was analyzed following digestion with endonuclease *Cla*I, and a correct one was identified and designated pFP200.

5 Next DNA sequences encoding six consecutive histidine residues were inserted into pFP200. Such sequences were carried on a synthetic double stranded oligonucleotide (SF25/26) with the following sequence:

G S H H H H H H S R (SEQ ID NO:10)

10 5'HO-GATCCCATCACCATCACCATCACTCTA (SEQ ID NO:11)

GGTAGTGGTAGTGGTAGTGGAGATCTAG-OH 5' (SEQ ID NO:12)

The amino acid sequence encoded by this oligonucleotide when it is inserted in the correct orientation into the *Bam*HI site of pFP200 is shown in 15 one-letter code above the DNA sequence. DNA of pFP200 was digested with endonuclease *Bam*HI and recovered by the GENECLEAN® procedure (Biol01, Inc., P.O. Box 2284, La Jolla, CA). An aliquot of this digested DNA (approximately 0.02 pmoles) was mixed with 20 oligonucleotide SF25/26 (10 pmoles), the 5' termini of which had not been phosphorylated. After incubation under ligation conditions for 5 h at 4 °C and 20 min at 23 °C, an aliquot was used to transform *E. coli* BL21. Transformants were selected for kanamycin resistance and 25 plasmid DNA of individual transformants was analyzed following digestion with endonucleases *Eco*RI and *Bam*HI. A correct plasmid was identified by the presence in the digest of a DNA band indicative of restoration of the *Bam*HI site at the promoter-proximal end of the oligo- 30 nucleotide sequence, resulting from insertion in the desired orientation. This plasmid was designated pFP202. Correct insertion of the oligonucleotide was verified by direct DNA sequencing as described above.

The plasmid vector pFP204 was constructed in an 35 analogous manner, by inserting into pFP200 a synthetic

double stranded oligonucleotide (SF29/30) with the following sequence:

G S H H H H H H (SEQ ID NO:13)

5'HO-GATCCCATCACCATCACCATCACTAAA (SEQ ID NO:14)

5 GGTAGTGGTAGTGGTAGTGATTCTAG-OH 5' (SEQ ID NO:15)

This oligonucleotide places a termination codon immediately following the six tandem His residues.

DP-1A.9 strains:

Next sequences encoding DP-1A were inserted into pFP202 at the BamHI site located between the T7 promoter and sequences encoding the His6 oligomer. DNA of plasmids pFP534 (encoding 101 aa DP-1A), pFP538 (encoding 2 repeats of 101 aa DP-1A), and pFP541 (8 repeats of 101 aa DP-1A) were digested with endonucleases BamHI and BglII, and pFP546 (16 repeats of 101 aa DP-1) was digested with BamHI, BglII, and EcoRI. The digests were fractionated by electrophoresis in low-melting agarose, and the ethidium bromide-stained band carrying the DP-1-encoding sequences was identified by size and excised. The excised gel bands were melted, and to each was added an aliquot of pFP202 DNA that had been digested with endonuclease BamHI. DNA was recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA) and incubated under ligation conditions for 2 h at 4 °C, followed by 20 min at 23 °C. An aliquot of ligated DNA was used to transform *E. coli* BL21(DE3), and transformants were selected for resistance to kanamycin.

Individual transformants were patched onto a sheet of cellulose acetate on the surface of LB agar containing kanamycin. After overnight growth, the cellulose acetate was transferred to a second plate on which a sheet of nitrocellulose had been placed on the surface of LB agar containing 1mM IPTG. After incubation for 3 h at 37 °C, the nitrocellulose sheet

was removed from under the cellulose acetate, blocked with "Blotto", and developed by immunochemical staining with anti-DP-1 serum as described below. Positive transformants, identified by blue color in this colony immunoassay, were picked from a replica master plate that had been inoculated at the same time as the immunoassay plate, with the same transformant colonies. The correct structure of plasmid DNA from positive transformants was verified following digestion with endonucleases BamHI and BglIII. Transformants in which the DP-1-encoding insert was inserted backwards (as identified by the formation of appropriately sized bands in the digest) gave a positive reaction on colony immunoassay, but the color yield was markedly less intense than those in the correct orientation. Transformants containing plasmids with correctly oriented inserts were identified and designated FP3211 (1 repeat of 101 aa), FP3217 (2 repeats), FP3203 (8 repeats) and FP3206 (16 repeats). DP-1 protein produced by strains FP3217, FP3203, and FP3206 was assayed by Western blot analysis as described below. All were shown to produce full-length protein of the expected size, detected by anti-DP-1 serum. In addition, a regular array of anti-DP-1-staining protein bands was observed, mainly at higher gel mobilities.

DP-1B.9 strains:

E. coli strains for the production of DP-1B.9 were constructed in a similar fashion by transferring DNA fragments encoding DP-1B.9 (SEQ ID NO.:81) (derived by digestion with BamHI and BglII of plasmids pFP156 and pFP158, containing 8 and 16 repeats of the 303 bp DNA monomer, respectively) into plasmid pFP202. The resulting production strains were designated FP2121 (8repeats) and FP2123 (16 repeats). Both strains were

shown by Western Blot analysis to produce full-length protein of the expected size.

DP-1B.16 strains:

E. coli strains for the production of DP-1B.16 (SEQ ID NO.:82) were constructed in a similar fashion by transferring DNA fragments encoding DP-1B.16 (derived by digestion with BamHI and BglII of plasmids pFP662 and pFP665 containing 8 and 16 repeats of the 303 bp DNA monomer, respectively) into plasmid pFP204. The resulting production strains were designated FP3350 (8 repeats) and FP3356 (16 repeats). Both strains were shown by Western Blot analysis to produce full-length protein of the expected size. Host cell FP3350 has been deposited with the ATCC under the terms of the Budapest Treaty and is identified by the ATCC number ATCC 69328 (deposited 15 June 1993).

DP-2A strains:

E. coli strains for the production of DP-2A were constructed in a similar fashion by transferring DNA fragments encoding DP-2A (derived by digestion with BamHI and BglII of plasmids pFP597 and pFP598, containing 8 and 16 repeats of the 357 bp DNA monomer, respectively) into plasmid pFP204. The resulting production strains were designated FP3276 (8 repeats) and FP3284 (16 repeats). Both strains were shown by Western Blot analysis to produce full-length protein of the expected size.

EXAMPLE 5

LARGE SCALE PRODUCTION, PURIFICATION AND QUANTITATION OF RECOMBINANT SILK VARIANT PROTEINS
Purification of DP-1A.9 (SEQ ID NO.:80):

Strain FP3203 was grown at 36 °C in a Fermgen fermenter (New Brunswick Scientific, New Brunswick, NJ) in 10 l of a medium containing:

$(\text{NH}_4)_2\text{SO}_4$

3.0 g

MgSO ₄	4.5 g
Na citrate · 2H ₂ O	0.47 g
FeSO ₄ · 7H ₂ O	0.25 g
CaCl ₂ · 2H ₂ O	0.26 g
Thiamine-HCl	0.6 g
Casamino acids	200 g
Biotin	0.05 g
K ₂ HPO ₄	19.5 g
NaH ₂ PO ₄	9.0 g
Glycerol	100 g
L-Alanine	10.0 g
Glycine	10.0 g
Glucose	200 g
PPG	5 mL
ZnSO ₄ · 7H ₂ O	0.08 g
CuSO ₄ · 5H ₂ O	0.03 g
MnSO ₄ · H ₂ O	0.025 g
H ₃ BO ₃	0.0015 g
(NH ₄) _n MO _x	0.001 g
CoCl ₂ · 6H ₂ O	0.0006 g

The fermenter was inoculated with 500 mL overnight culture of FP3203 in the same medium. The pH was maintained at 6.8 by addition of 5 N NaOH or 20% H₃PO₄. Dissolved O₂ was maintained at approximately 50%. When 5 the absorption at 600 nm had reached 10-15, production of DP-1 was induced by adding 5-g IPTG. After 3 h, cells were harvested by centrifugation and frozen. The yield was 314 g cell paste. Thawed cells (100 g paste) were suspended in 1000 mL buffer 8.0G containing 6 M 10 guanidine-HCl, 0.1 M NaH₂PO₄, 0.01 M Tris-HCl, 5 mM 2-mercaptoethanol, pH adjusted to 8.0 with NaOH. After stirring for 1 h at 23 °C, the lysate was clarified by centrifugation at 10,000 x g for 30 min, and the supernatant was filtered through Whatman No. 3 paper. 15 To the filtrate was added 200 mL packed volume of

Ni-nitrilotriacetic acid (NTA)-agarose (Qiagen, Inc.), which had been equilibrated with buffer 8.0G, recovered by filtration, and drained. The lysate-resin slurry was stirred at 23 °C for 24 h, then the resin was recovered 5 by filtration on Whatman No. 3 paper. The drained resin was suspended in 500 mL buffer 8.0G and packed into a chromatography column (5 cm diameter). The column was washed with 500 mL buffer 8.0G, then with successive 320 mL volumes of buffers of the same composition as 10 buffer 8.0G, but with the pH adjusted with NaOH to the following values: pH 6.3, 6.1, 5.9, 5.7, and 5.5. Effluent fractions of 40 mL were collected. DP-1 protein was located by immunoassay, as described above. Positive fractions were pooled and the pH was adjusted 15 to 8.0 with NaOH. Immunoassay and Western blot analysis revealed that approximately 50% of the material containing DP-1 sequences was adsorbed to the resin and recovered in the pooled fractions. The remaining material apparently lacks the C-terminal oligo-histidine 20 affinity tail, presumably as a result of premature termination of protein synthesis.

The concentration of 2-mercaptoethanol was adjusted to 17 mM, and the pooled material was stirred for 5 h at 23 °C. This material was reapplied to the same 25 Ni-NTA-agarose column, which had been re-equilibrated with buffer 8.0G. The column was then washed with 200 mL buffer 8.0G and 400 mL of buffer with a similar composition, but with a pH of 6.5, followed by 400 mL of a buffer composed of 0.1 M acetic acid adjusted to pH 30 6.5 with triethylamine, plus 5 mM 2-mercaptoethanol. DP-1 protein was eluted with 800 mL of a buffer composed of 0.1 M acetic acid adjusted to pH 5.0 with triethylamine, while 40 mL eluant fractions were collected. DP-1 protein was located by immunoassay. Positive 35 fractions were pooled and the buffer was removed by

lyophilization. Yield of lyophilized material was 100 mg, representing approximately 1% of the total protein present in the 100 g cell paste from which it was derived.

5 Amino acid analysis of the purified DP-1 is shown in Table I and is consistent with the predicted amino acid sequence, with impurities (as proteins of amino acid composition reflecting the overall composition of *E. coli* (Schaechter, M. et al., in *Escherichia coli and 10 Salmonella typhimurium*, Neidhardt, F. C. (ed) Washington D.C., American Association for Microbiology, p.5, (1987)) less than 7%.

TABLE I
Amino Acid Analysis DP1-A, 8-mer,
Recovered from FP3203

Amino Acid	Residues per Molecule		n Moles Experimental (Raw)
	Theoretical	Experimental	
Gly	383	367	10.91
Ala	235	[235]	6.98
Glx	92	98	2.91
Leu	40	40	1.32
Ser	37	37	1.09
Tyr	24	25	0.75
Arg	18	22	0.66
Met	3	3	0.09
His	6	8.7	0.26
Asx	0	6	0.18
Thr	1	4	0.13
Val	0	4	0.13
Ile	0	3	0.10
Phe	0	0	
Lys	0	3	0.10
Pro	0	0	0.00

Purity: 93%

Purification of DP-1B.16 (SEQ ID NO.:82):

Strain FP3350 was grown in 5 liters under conditions noted above. Thawed cell paste (154 g) was suspended in 1000 mL buffer 8.0G and stirred for 2 h at 23 °C. The lysate was clarified by centrifugation for 30 min at 10,000 x g. To the supernatant was added 300 mL (packed volume) of Ni-NTA agarose equilibrated with buffer 8.0gG. The mixture was stirred at 23 °C for 18 h, then the resin was recovered by centrifugation at 1,000 x g for 30 min. The resin was diluted to 800 mL with buffer 8.0G, mixed, and allowed to settle. Supernatant was removed and the settling procedure was repeated. The settled resin was then diluted with an equal volume of buffer 8.0G and packed into a chromatography column (5 cm diameter). The column was washed successively with (a) 1300 mL buffer 8.0G, (b) 500 mL buffer 8.0G containing 8 mM imidazole, (c) 100 mL buffer 8.0G, and (d) 500 mL buffer 6.5G (same composition as buffer 8.0G, but with the pH adjusted to 6.5 with NaOH). DP-1B.16 protein was finally eluted with buffer 5.5G (same composition as buffer 8.0G, but with the pH adjusted to 5.5 with NaOH). Fractions containing DP-1B.16 were identified by spot immunoassay, pooled, and concentrated approximately 40-fold by ultrafiltration using Centriprep 30 centrifugal concentrators (Amicon). Protein was precipitated by the addition of 5 volumes of methanol, incubating 16 h at 4 °C, recovered by centrifugation, washed twice with methanol and vacuum dried.

The yield of dried material was 287 mg, representing approximately 2% of the total protein present in the 154 g cell paste from which it was derived. Amino acid analysis is shown in Table II and is consistent with the predicted amino acid sequence, with impurities (as proteins of amino acid composition

reflecting the overall composition of *E. coli*) representing approximately 21% of the total protein in the sample.

TABLE II
Amino Acid Analysis
DP-1B16 8-mer Recovered from FP3350

Amino Acid	Residues per Molecule		nMoles Experimental (Raw)
	Theoretical	Experimental	
Gly	383	338	26.27
Ala	235	[235]	18.25
Glx	92	105	8.13
Leu	40	54	4.22
Ser	37	32	2.44
Tyr	24	25	1.95
Arg	18	30	2.32
Met	3	4.2	0.32
His	6	24.2	1.88
Asx	0	19.2	1.49
Thr	1	9.4	0.73
Val	0	13.5	1.05
Ile	0	10.7	0.83
Phe	0	7.3	0.57
Lys	0	10.1	0.78
Pro	0	8.6	0.67

Purity: 79%

Purification of DP-2A (SEQ ID NO.:83):

5 Strain FP3276 was grown in 5 liters under conditions noted above, except that the growth medium was supplements at inoculation with 0.375 g/l L-proline, and at induction with 0.1 g/l glycine and L-alanine and 0.0375 g/l L-proline. Thawed cell paste from two such 10 fermentations (150 g and 140 g, respectively) was suspended in 1000 mL each buffer 8.0G and stirred for 1 h at 23 °C. The lysate was clarified by centrifugation for 30 min at 10,000 x g. The supernatants were

combined and mixed with 300 mL (packed volume) of Ni-NTA agarose equilibrated with buffer 8.0G. The mixture was stirred at 23 °C for 18 h, then the resin was recovered by centrifugation at 1,000 x g for 30 min. The resin 5 was diluted to 800 mL with buffer 8.0G, mixed, and allowed to settle. Supernatant was removed and the settling procedure was repeated twice. The settled resin was then diluted with an equal volume of buffer 8.0G and packed into a chromatography column (5 cm 10 diameter). The column was washed successively with (a) 1350 mL buffer 8.0G, (b) 400 mL buffer 8.0G containing 8 mM imidazole, (c) 100 mL buffer 8.0G, and (d) 750 mL buffer 6.5G. DP-2A protein was finally eluted with buffer 5.5G. Fractions containing DP-1B.16 were 15 identified by spot immunoassay and pooled.

Of a total of 240 mL pooled fractions, 150 was removed and concentrated approximately 40-fold by ultrafiltration using Centriprep 30 centrifugal concentrators (Amicon). Protein was precipitated by the 20 addition of 5 volumes of methanol, incubating 16 h at 4 °C, recovered by centrifugation, washed twice with methanol and vacuum dried. The yield of dried material was 390 mg.

The remaining 90 mL pooled column fractions was 25 concentrated 8-fold using Centriprep 30 concentrators, diluted to the original volume with water and concentrated again. This procedure was repeated three additional times in order to remove guanidine to less than 5 mM. The material was finally lyophilized. The 30 weight of lyophilized material was 160 mg. Thus the total yield of purified DP-2A was 550 mg, representing approximately 2% of the total protein present in the 290 g cell paste from which it was derived.

Amino acid analysis of a sample of the lyophilized 35 material is shown in Table III and is consistent with

the predicted amino acid sequence, with impurities (as proteins of amino acid composition reflecting the overall composition of *E. coli*) representing less than 4% of the total protein in the sample.

TABLE III
Amino Acid Analysis
DP-2A, 8-mer Recovered from Strain FP3276

Amino Acid	Residues per Molecule		nMoles Experimental (Raw)
	Theoretical	Experimental	
Gly	373	351	16.98
Ala	185	[185]	8.95
Pro	169	158	7.64
Glx	130	93	4.51
Ser	51	48	2.35
Tyr	56	57	2.76
Met	3	2.0	0.10
His	6	9.2	0.45
Leu	1	1.8	0.09
Asx	0	ND	ND
Thr	1	ND	ND
Val	0	5.5	0.27
Ile	0	0	0.00
Phe	0	2.8	0.13
Lys	0	1.9	0.09
Arg	1	0	0.00

Purity: 96%

5 The present invention discloses the construction of several specific expression systems useful for the production of spider silk variant proteins. In order to leave no doubt that one of skill in the art might be able to use the elements of the instant invention to

10 produce the myriad of other spider silk variant proteins not specifically discussed, *E. coli* bacteria transformed with an expression vector (pFP204) devoid of synthetic spider silk variant DNA has been deposited with the ATCC

under the terms of the Budapest treaty and is identified by the ATCC number ATCC 69326. The expression pFP204 contained in the host cell *E. coli* HB101 comprises all the necessary restriction sites needed to clone 5 synthetic spider silk DNA of the instant invention and may be used to express any spider silk variant protein. In addition, the expression host strain *E. coli* BL21 (DE3) transformed with a plasmid pFP674 carrying DP-1B.16 coding sequences (SEQ ID NO.:82), has been 10 deposited with the ATCC under the terms of the Budapest treaty and is identified by the ATCC number ATCC 69328. This strain can be used to produce DP-1B according to this invention, or cured of plasmid by methods well known to those skilled in the art and transformed with 15 other expression vectors derived from pFP204.

EXAMPLE 6

SYNTHESIS AND EXPRESSION OF DP-1

ANALOG IN BACILLUS SUBTILIS

For expression in *Bacillus subtilis*, a DP-1 analog- 20 encoding gene from plasmid pFP141 was placed in a plasmid vector capable of replication in *B. subtilis*. DP-1 coding sequences were operably linked to a promoter derived from the levansucrase (*lvs*) gene of *Bacillus amyloliquefaciens* in such a manner that the N-terminal 25 amino acid sequence coded by the levansucrase gene, which comprises a secretion signal sequence, was fused to the DP-1 sequence at its N-terminus. Gene fusions of this type have been shown, in some cases, to promote the production and secretion into the extracellular medium 30 of foreign proteins (Nagarajan et al. U.S. Patent 4,801,537).

As illustrated in Fig. 15, to prepare the DP-1 analog gene for transfer into the appropriate vector for *B. subtilis*, the endonuclease *Bgl*II site at the proximal 35 end of the DP-1 coding sequence in plasmid pFP541 was

first converted to an EcoRV site by inserting a synthetic oligonucleotide. DNA of plasmid pFP541 was digested with endonuclease BglII. Approximately 0.1 pmole of the linearized plasmid DNA was then 5 incubated under ligation conditions with 10 pmoles of a synthetic double stranded oligonucleotide (SI9/10) with the following sequence:

5'HO-GATCAGATATCG (SEQ ID NO:16)
TCTATAGCCTAG-OH 5' (SEQ ID NO:17)

10 Ampicillin resistant transformants of *E. coli* HMS174 were screened for plasmid DNA containing an EcoRV site provided by the synthetic oligonucleotide sequence. A plasmid containing an EcoRV site was identified and designated pFP169b (Figure 15). Next the DNA fragment 15 carrying DP-1 coding sequences was isolated from pFP169b following digestion with endonucleases EcoRV and BamHI and separation of the resulting DNA fragments by agarose gel electrophoresis. A band of the appropriate size was excised from the ethidium bromide stained gel and DNA 20 was recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA).

The plasmid vector pBE346 contains replication origins that confer autonomous replication in both *E. coli* and *B. subtilis*, as well as antibiotic 25 resistance markers selectable in *E. coli* (ampicillin) and *B. subtilis* (kanamycin). In addition, the plasmid contains the *lvs* promoter and secretion signal operably linked to a staphylococcal protein A gene. The protein A gene is bounded by an EcoRV site at its proximal end, 30 separating it from the *lvs* signal sequence, and a BamHI site at its distal end. The complete DNA sequence of pBE346 (Figure 14) is shown in SEQ ID NO.:79 and in Figure 14. In order to remove the protein A gene and allow for its replacement by the DP-1 gene, DNA of 35 plasmid pBE346 was digested with endonucleases EcoRV and

BamHI and the appropriate sized fragment was isolated following agarose gel electrophoresis. DNA was recovered from the ethidium bromide stained gel band by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, 5 La Jolla, CA).

DNA fragment purified from pFP169b (above) was mixed with the DNA fragment purified from pBE346 and incubated under ligation conditions. Ligated DNA was used to transform *E. coli* HMS174, and ampicillin 10 resistant transformants were screened by examining plasmid DNA for the presence of appropriately sized fragments following digestion with endonucleases EcoRV and BamHI. A correct plasmid was identified and designated pFP191 (Figure 15).

15 DNA of plasmid pFP191 was used to transform competent cells of *B. subtilis* BE3010 (*trp lys apr npr sacB*). Transformants were selected for resistance to kanamycin. BE3010 was derived from *B. subtilis* BE1500, 20 (*trpC2, metB10, lys3, delta-aprE, delta-npr, sacB::ermC*) which has been described by Nagarajan et al., *Gene*, 114, 121, (1992) by transforming competent BE1500 cells with DNA from *B. subtilis* 1S53 (Bacillus Genetic Stock Center, Ohio State University) and selecting for methionine prototrophs. Transformation of competent 25 cells was carried out essentially as described by Nagarajan et al., U.S. Patent 4,801,537.

Kanamycin resistant transformants of BE3010 were screened for the ability to produce DP-1 by colony immunoassay. Colonies were grown on a cellulose acetate 30 disk placed on the surface of a plate containing TBAB agar plus 5 micrograms per mL kanamycin. After colonies had developed at 37 °C, the cellulose acetate disk was transferred to a fresh plate containing the same medium plus 0.8% sucrose, and placed over a nitrocellulose disk 35 which was placed on the surface of the agar. After

incubation for 3 h at 37 °C, the nitrocellulose disk was removed and stained with anti-DP-1 serum, peroxidase-conjugated goat anti-rabbit IgG, and 4-chloro-1-naphthol plus hydrogen peroxide as described above. Positively 5 staining images of the colonies were observed, indicating the production and excretion of DP-1, compared to a negative control strain containing a plasmid with no DP-1 coding sequences. The positive 10 strain was designated FP2193. FP2193 has been deposited with the ATCC under the terms of the Budapest Treaty and is identified by the ATCC number, ATCC 69327.

The production and excretion of DP-1 by FP2193 was assayed in liquid culture. Strain FP2193 was grown in Medium B, containing, per liter, 33 g Bacto-tryptone 15 (Difco), 20 g yeast extract, 7.4 g NaCl, 12 mL 3N NaOH, 0.8 g Na₂HPO₄, 0.4 g KH₂PO₄, 0.2% casamino acids (Difco), 0.5% glycerol, 0.06 mM MnCl₂, 0.5 nM FeCl₃, pH 7.5. After growth for 3.5 h at 37 °C, production of DP-1 was induced by the addition of sucrose to 0.8%. After 4 h 20 additional incubation at 37 °C, a sample of 0.5 mL was analyzed. Cells were removed by centrifugation. The upper 0.4 mL of supernatant was removed and phenyl-methane sulfonyl fluoride (PMSF) was added to 2 mM. The residual supernatant was removed and discarded. The 25 cell pellet was suspended in 0.32 mL 50mM EDTA, pH8.0, and lysed by the addition of 0.08 mL 10 mg/mL egg white lysozyme in the same buffer, plus 2mM PMSF. After incubation for 60 min at 37 °C, 0.01 mL 2M MgCl₂ and 0.001 mL 1 mg/mL deoxyribonuclease I were added, and 30 incubation continued for 5 min at 37 °C. Aliquots (5 microliters) of each fraction, cell lysate and supernatant, were analyzed by SDS gel electrophoresis and electroblotting as described above. The blot was stained with anti-DP-1 serum. Several positively 35 staining bands were observed in the supernatant

fraction, and only a trace of positive band in the cell lysate. The host strain BE3010 containing no DP-1 coding DNA sequences produced no positively staining bands. Thus *B. subtilis* strain FP2193 was shown to 5 produce DP-1 analog protein and to excrete it efficiently into the extracellular medium.

EXAMPLE 7

DP-1B Production in *Pichia pastoris*

1. Synthetic Gene DP-1B.33

10 A set of genes encoding DP-1B, designated DP-1B.33, were designed to encode proteins of the same repeating sequence as DP-1B.9 and DP-1B.16, but to use predominantly codons favored in the highly expressed alcohol oxidase genes of *Pichia pastoris*.

15 a. Oligonucleotides

Synthetic genes encoding DP-1B.33 were assembled from four double stranded synthetic oligonucleotides whose sequences are shown in Figure 16. The oligonucleotides were provided by the manufacturer (Midland 20 Certified Reagents, Midland, TX) in single-stranded form with 5'-OH groups not phosphorylated. For annealing to the double-stranded form, complementary single stranded oligonucleotides (667 pmoles each) were mixed in 0.2 ml buffer containing 0.01 M Tris-HCl, 0.01 M MgCl₂, 0.05 M 25 NaCl, 0.001 M dithiothreitol, pH 7.9. The mixture was heated in boiling water for 1 min, then allowed to cool slowly to 23 °C over approximately 3 h.

The four double-stranded oligonucleotides were separately cloned by inserting them into a plasmid 30 vector pFP206. DNA of plasmid pFP206 was digested with endonucleases BamHI and BglII and purified by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). To approximately 0.1 pmole of the eluted plasmid DNA was added 10 pmoles of one of the double- 35 stranded oligonucleotides P1, P2, P3, or P4. The four

plasmid-oligonucleotide mixtures were incubated under ligation conditions for 20 h at 4 °C, then ligation was terminated by incubation for 2 min at 70 °C. Ligated DNA was then digested with endonuclease HindIII to 5 linearize any remaining parental pFP206. Aliquots of ligated DNA were used to transform *E. coli* HB101 and ampicillin resistant transformants were selected. Clones containing oligonucleotides P1, P2, P3, or P4 were identified by screening plasmid DNA isolated from 10 individual transformants with endonucleases BamHI and PstI. In plasmids with inserts in the desired orientation, the shorter of two BamHI-PstI fragments of pFP206 is lengthened by the length of the cloned oligonucleotide. Plasmid DNA from putative clones was 15 further characterized by digestion with endonucleases BamHI and BglII and analysis by electrophoresis in 3.8% MetaPhor agarose (FMC) to verify that the plasmid had acquired a single copy of the oligonucleotide in the correct orientation. Correct clones were identified and 20 their plasmids were designated pFP685 (oligonucleotide P1, SEQ ID NOS.:84, 85, and 86), pFP690 (oligonucleotide P2, SEQ ID NOS.:87, 88, and 89), pFP701 (oligonucleotide P3, SEQ ID NOS.:90, 91, and 92), and pFP693 (oligo- 25 nucleotide P4, SEQ ID NOS.:93, 94, and 95). Sequences of all four cloned oligonucleotides were verified by DNA sequencing.

b. Assembly of the gene

For assembly of subsequence P1,P2, plasmid pFP685 (P1, SEQ ID NOS.:84, 85, and 86) was digested with 30 endonucleases PstI and BamHI, and plasmid pFP690 (P2, SEQ ID NOS.:87, 88, and 89) was digested with endonucleases PstI and BglIII. Digested plasmid DNA was fractionated by electrophoresis in a 1.2% agarose (low melting, BioRad, Hercules, CA) gel. Ethidium bromide- 35 stained bands containing the oligonucleotide sequences,

identified by their relative sizes, were excised, the excised bands combined, and the DNA recovered from melted agarose by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined 5 DNA fragments were incubated under ligation conditions and an aliquot was used to transform *E. coli* HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated from several transformants, digested with endonucleases BamHI and BglIII, and 10 analyzed by agarose gel electrophoresis. Plasmid containing insert of the expected size was identified and designated pFP707.

Assembly of subsequence P3,P4 was accomplished in the same manner as the subsequence P1,P2, starting, 15 however, with plasmids pFP701 (digested with PstI and BamHI) and pFP693 (digested with PstI and BglIII). Plasmid containing the P3,P4 subsequence was identified and designated pFP709.

For assembly of the DNA monomer (P1,P2,P3,P4), 20 plasmid pFP707 (P1, P2) was digested with endonucleases PstI and BamHI, and plasmid pFP709 (P3,P4) was digested with endonucleases PstI and BglIII. Digested plasmid DNA was fractionated by electrophoresis in a 1.2% low melting agarose gel. Ethidium bromide-stained bands 25 containing the P1,P2 and P3,P4 sequences, respectively, identified by their relative sizes, were excised, the excised bands combined, and the DNA recovered from melted agarose by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The eluted combined 30 DNA fragments were incubated under ligation conditions and an aliquot was used to transform *E. coli* HB101. Ampicillin-resistant transformants were selected. Plasmid DNA was isolated from several transformants, digested with endonucleases BamHI and BglIII, and 35 analyzed by agarose gel electrophoresis. Plasmid

containing an insert of the expected size was identified and designated pFP711. The DNA insert in plasmid pFP711 was verified by direct DNA sequencing.

c. Polymerization of the gene

5 The synthetic gene was extended by sequential doubling, starting with the monomer sequence in pFP711. For doubling any insert sequence, an aliquot of plasmid DNA was digested with endonucleases PstI and BamHI, and a separate aliquot of the same plasmid was digested with 10 endonucleases PstI and BglII. Digests were fractionated by electrophoresis on low melting agarose (BioRad, CA), and ethidium bromide stained fragments containing insert sequences were identified by their relative sizes. The two insert-containing fragments, purified by 15 electrophoresis and recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA), were combined and incubated under ligation conditions. At the third doubling, the two fragments in the BamHI digest were not adequately separated, so the eluted band 20 contained both fragments. In this case a two-fold excess of the BglII-PstI fragment was used in the ligation. An aliquot of the ligated DNA was used to transform *E. coli* HB101. Ampicillin resistant transformants were selected. Plasmid DNA was isolated 25 from several transformants, digested with endonucleases BamHI and BglII, and analyzed by agarose gel electrophoresis. Plasmid containing an insert of the expected size was identified.

By this procedure a series of plasmids was 30 constructed containing 2, 4, 8, and 16 tandem repeats of the DNA monomer sequence P1,P2,P3,P4, encoding the series of DP-1B.16 analogs. These plasmids were designated pFP713 (2 repeats), pFP715 (4 repeats), pFP717 (8 repeats), and pFP719 (16 repeats), and p723 35 (16 repeats), respectively.

2. Expression of DP-1 and DP-2 analog genes in *Pichia pastoris*

a. Growth and Assays

For the growth of cultures to assess production levels, 20 ml BMGY (per liter: 13.4 g yeast nitrogen base with ammonium sulfate (Difco), 10 g yeast extract, 20 g peptone, 0.4 mg biotin, 100 ml 1 M potassium phosphate buffer, pH 6.0, 10 ml glycerol) in a 125 ml baffled Erlenmeyer flask was inoculated at an absorption 10 (A_{600 nm}) of approximately 0.1 with cells eluted from a YPD agar plate (containing per liter: 10 g yeast extract (Difco), 20 g peptone, 20 g Bacto agar (Difco), 20 g D-glucose), which had been grown 2 days at 30 °C. The culture was shaken at 30 °C until the A_{600 nm} reached 15 approximately 25 (2 days), at which time cells were harvested by centrifugation (5 min at 1500 x g). Supernatant was discarded and the cells resuspended in 6 ml BMMY (same as BMGY, except with 5 ml methanol per liter in place of glycerol). The culture was shaken at 20 30 °C, and 0.005 ml methanol per ml culture was added every 24 h. Samples (1 ml) were taken immediately after resuspension and at intervals. Cells were immediately recovered by centrifugation in a microfuge (2 min at 6000 x g). Where secretion was to be assayed, the top 25 0.7 ml supernatant was removed and frozen in dry ice ("culture supernatant" fraction). The drained cell pellet was frozen in dry ice and stored at -70 °C.

Cells were lysed by shaking with glass beads. The thawed pellet was washed with 1 ml cold breaking buffer 30 (50 mM sodium phosphate, pH 7.4, 1 mM EDTA, 5% (v/v) glycerol, 1 mM phenyl methane sulfonyl flouride), and resuspended in 0.1 ml of the same buffer. Glass beads (acid washed, 425-600 microns; Sigma Chemical Co.) were added until only a meniscus was visible above the beads, 35 and the tubes subjected to mixing on a vortex type mixer

for two intervals of 4 min, cooling on ice between. Cell breakage was verified by microscopic examination. After complete breakage, 0.5 ml breaking buffer was added and mixed. Debris and beads were pelleted in the 5 microfuge (10 min), and 0.5 ml supernatant (soluble cell extract) removed. The debris was then extracted twice with additional 0.5 ml portions of breaking buffer, and the 0.5 ml supernatants combined with the first extract ("soluble cell extract" fraction). The 10 debris was then extracted three times with 0.5 ml portions of buffer 6.5G, containing 0.1 M sodium phosphate, 0.01 M Tris-HCl, 6M guanidine-HCl, pH 6.5. The combined supernatants comprised the "insoluble cell extract" fraction.

15 For analysis by polyacrylamide gel electrophoresis, extracts were diluted approximately 1000-fold into sample preparation buffer (0.0625 M Tris-HCl, pH 6.8, 2% w/v Na-dodecyl sulfate, 0.0025% w/v bromphenol blue, 10% v/v glycerol, 2.5% v/v 2-mercaptoethanol), and incubated 20 in a boiling water bath for 5 min. Aliquots (5-15 μ l) were applied to an 8% polyacrylamide gel (Novex) and subjected to electrophoresis until the dye front was less than 1 cm from the bottom of the gel. Protein bands were transferred electrophoretically to a sheet of 25 nitrocellulose, using an apparatus manufactured by Idea Scientific, Inc. The buffer for transfer contained (per liter) 3.03 g Trishydroxymethyl aminomethane, 14.4 g glycine, 0.1% w/v SDS, 25% v/v methanol.

The nitrocellulose blot was stained immuno- 30 chemically as follows. Protein binding sites on the sheet were blocked by incubation with "Blotto" (3% nonfat dry milk, 0.05% Tween 20, in Tris-saline (0.1 M Tris-HCl, pH 8.0, 0.9% w/v NaCl)) for 30 min at room temperature on a rocking platform. The blot was then 35 incubated for 1 h with anti DP-1 serum, diluted 1:1000

in "Blotto", washed with Tris saline, and incubated for 1 h with horseradish peroxidase-conjugated goat anti-rabbit IgG serum (Kierkegaard and Perry Laboratories, Gaithersburg, MD), diluted 1:1000 in "Blotto". After 5 again washing with Tris-saline, the blot was exposed to a solution of 18 mg 4-chloro-1-naphthol in 6 ml methanol, to which had been added 24 ml Tris-saline and 30 μ l 30% H_2O_2 .

For quantitation of DP-1 antigen levels in various 10 fractions, aliquots (1 μ l) of serial dilutions in buffer 6.5G were spotted onto nitrocellulose, along with various concentrations of a standard solution of purified DP-1 8-mer (8 repeats of 101 amino acid residues). The nitrocellulose sheet was then treated as 15 described above for the Western blot. The concentration of DP-1 antigen in each sample was estimated by matching the color intensity of one of the standard spots.

b. Production strains

(1) Vectors

20 To construct yeast strains for production of DP-1, cloned synthetic DP-1-coding DNA sequences were inserted into plasmid vectors which were derived from the plasmids pHIL-D4 (obtained from Phillips Petroleum Co.), or pPIC9 (obtained from Invitrogen Corp.). The 25 structure of pHIL-D4 is illustrated in Figure 17. The plasmid includes a replication origin active in *E. coli* (but not in yeast) and ampicillin and kanamycin resistance markers that are selectable in *E. coli*. The kanamycin resistance marker also confers resistance to 30 the antibiotic G418 in yeast. The plasmid includes regions homologous to both ends of the *Pichia pastoris* AOX1 gene. The upstream region includes the AOX1 promoter, expression from which is inducible by methanol. Sequences to be expressed are inserted 35 adjacent to the AOX1 promoter. Downstream are sequences

encoding the AOX1 polyadenylation site and transcription terminator, the kanamycin marker, and the *Pichia pastoris* HIS4 gene. In pHIL-D4 no translated sequences are provided upstream from the sequences to be 5 expressed. The vector pPIC9 (Figure 18) is similar to pHIL-D4, except it includes, adjacent to the AOX1 promoter, sequences encoding the signal sequence and pro- sequence of the *Saccharomyces cerevisiae* alpha-mating factor gene. Also, pPIC9 lacks the kanamycin 10 resistance gene of pHIL-D4.

A BamHI site in pPIC9, located immediately upstream of the 5' end of the alpha-mating factor gene was removed, and the sequences restored to those resembling the natural AOX1 gene, by polymerase chain reaction 15 (PCR) (Perkin Elmer Cetus, CA). Fragments of pPIC9 were amplified separately using the following primer pairs:

LB1: 5'-CAACTAATTATCGAAACGATGAGATTCC -3' (SEQ ID NO.:98)

LB6: 5'-CTGAGGAACAGTCATGTCTAAGG -3' (SEQ ID NO.:99)

20 and

LB2: 5'-GGAAATCTCATCGTTCGAATAATTAGTTG -3' (SEQ ID NO.:100)

LB5: 5'-GAAACGCAAATGGGGAAACAAACC -3' (SEQ ID NO.:101)

PCR reactions were carried out in a Perkin Elmer 25 Cetus DNA Thermal Cycler, using the Perkin Elmer Cetus GeneAmp kit with AmpliTaq[®] DNA polymerase. Instructions provided by the manufacturer were followed. The template DNA was approximately 0.2 ng pPIC9 DNA digested with endonucleases BglII and PvuII and subsequently 30 recovered by the GENECLEAN[®] procedure (Biol01, Inc., P.O. Box 2284, La Jolla, CA). The PCR program included (a) 1 min at 94 °C; (b) 4 cycles consisting of 1 min at 94 °C, 2 min at 45 °C, 1 min at 72 °C; (c) 25 cycles consisting of 1 min at 94 °C, 1 min at 60 °C, 1 min at 35 72 °C (extended by 10 sec each cycle); and (d) 7 min at 72 °C. Products were recovered from the two separate

PCR reactions by the GENECLEAN® procedure (P.O. Box 2284, La Jolla, CA) and mixed in approximately equimolar amounts. This mixture was used as template for a second round of PCR using primers LB5 and LB6. For this 5 reaction, the PCR program included (a) 1 min at 94 °C; (b) 25 cycles consisting of 1 min at 90 °C, 1 min at 60 °C, 1 min at 72 °C (extended 10 sec per cycle); and (d) 7 min at 72 °C. The PCR product was recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, 10 La Jolla, CA), then digested with endonucleases NsiI and EcoRI and again recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The fragment was purified by electrophoresis in 1.5% low melting agarose (BioRad). DNA was recovered from the 15 excised gel band by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). This fragment was substituted for the analogous fragment in pPIC9. For this purpose, pPIC9 was digested with endonucleases NsiI and EcoRI. The larger fragment was purified by 20 electrophoresis in a 1.2% low melting agarose gel and recovered from the excised gel band by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA). The PCR fragment and the large pPIC9 fragment were ligated under standard conditions, and the ligation was 25 used to transform *E. coli* HB101. Ampicillin resistant transformants containing the correct plasmid were identified by screening plasmid DNA for the absence of the BamHI site. The correct plasmid was designated pFP734. The DNA sequence of pFP734 in the affected 30 region, verified by DNA sequencing is shown in Figure 19 (SEQ ID NOS.:96 and 97).

DNA sequences encoding six consecutive histidine residues were inserted into pHIL-D4. Such sequences were carried on a synthetic double stranded oligo- 35 nucleotide (SF47/48) with the following sequence:

M G S H H H H H H H End	SEQ ID NO.:102
5' HO-AATTATGGGATCCCATCACCATCACCATCACT	SEQ ID NO.:103
TACCCCTAGGGTAGTGGTAGTGGTAGTGGATTAA-OH 5'	SEQ ID NO.:104

5

The amino acid sequence encoded by this oligonucleotide when it is inserted in the correct orientation into the EcoRI site of pHIL-D4 is shown in one-letter code above the DNA sequence. DNA of pHILD4 was digested with endonuclease EcoRI and recovered by the GENECLEAN® procedure (Biol01, Inc., P.O. Box 2284, La Jolla, CA). An aliquot of this digested DNA (approximately 0.02 pmoles) was mixed with oligonucleotide SF47/48 (10 pmoles), the 5' termini of which had not been phosphorylated. After incubation under ligation conditions for 19 h at 4 °C, an aliquot was used to transform *E. coli* HB101. Transformants were selected for ampicillin resistance and plasmid DNA of individual transformants was analyzed following digestion with endonucleases PvuII and BamHI. A correct plasmid was identified by the presence in the digest of a DNA band indicative of the BamHI site at the promoter-proximal end of the oligonucleotide sequence, resulting from insertion in the desired orientation. This plasmid was designated pFP684. Correct insertion of the oligonucleotide was verified by direct DNA sequencing.

The plasmid vector pFP743 was constructed in an analogous manner, by substituting for sequences between NotI and EcoRI sites in pFP734 a synthetic double stranded oligonucleotide (SF55/56) with the following sequence:

F G S Q G A End	SEQ ID NO.:105
5' HO-AATTGGGATCCCAGGGTGCTTAA	SEQ ID NO.:106
35 GCCTAGGGTCCCACGAATTCCGG-OH 5'	SEQ ID NO.:107

DNA of pFP734 was digested with endonucleases NotI and EcoRI, then recovered by the GENECLEAN® procedure (Bio101, Inc., P.O. Box 2284, La Jolla, CA).

Oligonucleotide SF55/56 was inserted by ligation as 5 described above. A correct plasmid was identified by the presence of a new fragment upon digesting plasmid DNA with endonucleases BamHI and BglII, and designated pFP743. Correct oligonucleotide insertion was verified by direct DNA sequencing.

10 (2) DP-1B.33 strains

Next, sequences encoding DP-1B were inserted into pFP684 and pFP743 at the respective unique BamHI sites located between the AOX1 promoter and sequences encoding the His6 oligomer. DNA (approximately 2 micrograms) of 15 plasmids pFP717 (encoding 8 repeats of 101 aa DP-1B) and pFP719 (encoding 16 repeats of 101 aa DP-1B) were digested with endonuclease BamHI and BglII. The digests were fractionated by electrophoresis in low-melting agarose, and the ethidium bromide-stained band carrying 20 the DP-1B-encoding sequences was identified by size and excised. The excised gel bands were melted, and to each was added an aliquot of pFP684 or pFP743 DNA that had been digested with endonuclease BamHI. DNA was recovered by the GENECLEAN® procedure (Bio101, Inc., 25 P.O. Box 2284, La Jolla, CA) and incubated under ligation conditions for 3 h at 13 °C. An aliquot of ligated DNA was used to transform *E. coli* HB101, and transformants were selected for resistance to ampicillin.

30 Individual transformants were screened by digesting plasmid DNA with endonucleases BamHI and BglII. Correct plasmids were identified by the presence of a fragment of the expected size containing the DP-1B.33 gene. Plasmids derived from the vector pFP684 were designated 35 pFP728 (encoding 8 repeats of 101 amino acids DP-1B) and

pFP732 (encoding 16 repeats of 101 amino acids DP-1B).

Those derived from the vector pFP743 were designated pFP748 (encoding 8 repeats of 101 amino acids DP-1B) and pFP752 (encoding 16 repeats of 101 amino acids

5 DP-1B).

Each of these plasmids was used to transfer the DP-1B gene to *Pichia pastoris* strain GS115 (his4) by spheroplast transformation essentially according to Cregg et al. (Mol. Cell. Biol. 5, 3376-3385 (1985)).

10 The *Pichia* strain was grown in 200 ml YPD medium in a 500 ml baffled flask at 30 °C to A_{600nm} of 0.3 to 0.4. Cells were harvested by centrifugation at 1500 x g for 5 min at room temperature, then washed with 20 ml

15 sterile water, followed by 20 ml fresh SED (1 M sorbitol, 25 mM EDTA, pH 8.0, 50 mM DTT), and 20 ml 1 M sorbitol. Cells were resuspended in 20 ml SCE (1 M sorbitol, 1 mM EDTA, 10 mM sodium citrate, pH 5.8), and zymolyase (15 ml stock solution containing 3 mg/ml Yeast Lytic Enzyme from *Arthrobacter luteus* (ICN Corp.;

20 specific activity 100,000 u/g)) was added.

Spheroplasting was monitored by diluting 0.2 ml aliquots into 0.8 ml 5% SDS and measuring A_{600nm}. Digestion was continued until 70-80% spheroplasting was obtained.

25 Spheroplasts were then harvested by centrifugation at 750 x g for 10 min at room temperature, washed once with 10 ml 1 M sorbitol and once with 10 ml CAS (1 M sorbitol, 10 mM Tris-HCl, pH 7.5, 10 mM CaCl₂), and finally resuspended in 0.6 ml CAS. To 0.1 ml spheroplast suspension was added 1-5 micrograms linear 30 DNA fragments in CAS, prepared by digesting plasmid DNA with endonuclease BglII and recovering the fragments by the GENECLEAN® procedure (Biol01, Inc., P.O. Box 2284, La Jolla, CA). PEG solution (1 ml containing 20% w/v PEG 3350 (Fisher Scientific Co,) in 10 mM Tris-HCl, 35 pH 7.5, 10 mM CaCl₂) was added, mixed gently, and

incubated 10 min at room temperature. Spheroplasts were recovered by centrifugation as above. The drained pellet was resuspended in 0.15 ml SOS (1 M sorbitol, 0.3 vol/vol medium YPD, 10 mM CaCl₂, incubated at room temperature 20 min, and diluted with 0.85 ml 1 M sorbitol. Washed spheroplasts were mixed with 15 ml RD agarose (containing, per liter: 186 g sorbitol, 10 g agarose, 20 g D-glucose, 13.4 g yeast nitrogen base without amino acids (Difco), 0.4 mg biotin, 50 mg each 5 L-glutamic acid, L-methionine, L-lysine, L-leucine, L-isoleucine, and 20 ml 50x His assay medium. The composition of 50x His assay medium was as follows (per liter): 50 g D-glucose, 40 g sodium acetate, 6 g ammonium chloride, 0.4 g D,L-alanine, 0.48 g L-arginine-HCl, 0.8 g L-asparagine monohydrate, 0.2 g L-aspartic acid, 0.6 g L-glutamic acid, 0.2 g glycine, 0.2 g D,L-phenylalanine, 0.2 g L-proline, 0.1 g D,L-serine, 0.4 g D,L-threonine, 0.5 g D,L-valine, 20 mg adenine sulfate, 20 mg guanine hydrochloride, 20 mg uracil, 10 20 mg xanthine, 1 mg thiamine-HCl, 0.6 mg pyridoxine-HCl, 0.6 mg pyridoxamine-HCl, 0.6 mg pyridoxal-HCl, 1 mg Ca pantothenate, 2 mg riboflavin, 2 mg nicotinic acid, 0.2 mg para-aminobenzoic acid, 0.002 mg biotin, 0.002 mg folic acid, 12 g monopotassium phosphate, 12 g 15 dipotassium phosphate, 4 g magnesium sulfate, 20 mg ferrous sulfate, 4 mg manganese sulfate, 20 mg sodium chloride, 100 mg L-cystine, 80 mg D,L-tryptophane, 200 mg L-tyrosine. Spheroplasts in RD agarose (5 ml aliquots) were plated on RDB plates with the same 20 25 30 composition as RD, but with 20 g agar (Difco) per liter in place of agarose.

Plates were incubated at 30 °C for 3-4 days. Histidine prototrophic transformants were picked and patched onto MGY plates containing (per liter) 15 g 35 agar, 13.4 g yeast nitrogen base without amino acids,

0.4 mg biotin, 10 ml glycerol. Replicas were patched onto a sheet of cellulose acetate on the surface of MGY agar. After 2 days growth at 30 °C, the cellulose acetate was transferred to a second plate on which a 5 sheet of nitrocellulose had been placed on the surface of MM agar with the same composition as MGY except 0.5% v/v methanol instead of glycerol. After incubation for 1-3 days at 30 °C, the nitrocellulose sheet was removed from under the cellulose acetate, blocked with "Blotto", 10 and developed by immunochemical staining with anti-DP-1 serum as described above. Positive transformants, identified by blue color in this colony immunoassay, were picked from the MGY master plate. Transformants were also tested for growth on MM agar. DP-1 protein 15 produced by immunoassay positive strains was assayed by Western blot analysis as described above. Several were shown to produce full-length protein of the expected size, detected by anti-DP-1 serum.

(2) DP-1B Production

20 DP-1B production by two such transformants is illustrated in Figures 20 and 21. Figure 20 shows intracellular production, after various times of methanol induction, by strain YFP5028, which was derived by transforming *Pichia pastoris* GS115 with plasmid 25 pFP728. This strain produces DP-1B species of 5 different sizes, as indicated by Western blot analysis, consisting of 8, 11, 13, 15 and greater than 20 repeats of the 101-amino acid residue monomer, respectively. It was identified among *Pichia* transformants by its ability 30 to grow on YPD medium containing 0.5 mg/ml antibiotic G418, presumably indicative of the presence of multiple copies of the pFP728-derived insert. Total production of DP-1B was in excess of 1 g per liter culture. Figure 21 shows the intracellular and extracellular 35 production of DP-1B by strain YFP5093, which was derived

by transformation of *Pichia pastoris* GS115 with plasmid pFP748. A significant fraction of the DP-1B produced was recovered from the extracellular culture supernatant.

5

EXAMPLE 8

Demonstration of the Solutioning and Extrusion of Fibers from a Recombinantly Synthesized Analog to Spider Dragline Protein

For fiber spinning, DP-1B was purified by ion exchange chromatography. Frozen cell paste of *E. coli* FP3350 was thawed, suspended in 0.02 M Tris-HCl buffer, pH 8.0 (Buffer A), and lysed by passage through a Mantin-Gaulin homogenizer (3-4 passes). Cell debris was removed by centrifugation, and the soluble extract was heated to 60°C for 15-min. Insoluble material was again removed by centrifugation, and the soluble heat-treated extract was adjusted to pH-8.0 and diluted to conductivity less than 0.025-M applied to a column of SP-Sepharose Fast Flow (Pharmacia, Piscataway, NJ) equilibrated with Buffer A. The column was washed with Buffer A and eluted with a linear gradient from 0 to 0.5 M NaCl in Buffer A. DP-1B-containing fractions were identified by gel electrophoresis and immunoblotting as described above, pooled, and DP-1B was recovered by precipitation with 4 volumes of methanol at 0°C and centrifugation. Pellets were washed three times with methanol and dried in vacuum. This material was found to be greater than 95% pure DP-1B as determined by amino acid analysis.

Briefly, the process of producing useful fibers from purified DP-1 protein involves the steps of dissolution in HFIP, followed by spinning of the solution through a spinneret orifice to obtain fibers. Physical properties such as tenacity, elongation, and initial modulus were measured using methods and

instruments which conformed to ASTM Standard D 2101-82, except that the test specimen length was one inch. Five breaks per sample were made for each test.

Wet Spinning of Silk Fibers from HFIP Solution:

5 DP-1 was added to hexafluoroisopropanol (HFIP) in a polyethylene syringe to make a 20% solution of DP-1 in HFIP. The solution was mixed thoroughly, by pumping back and forth between two syringes and allowed to stand overnight.

10 The 20% solids solution of DP-1 in HFIP was transferred to a syringe fitted with a scintered stainless steel DYNALLOG® filter (X7). The syringe was capped and periodically vented to disengage air bubbles trapped in the solution. A syringe pump was then used 15 to force the solution through the filter and out of the syringe through a 5 mil diameter by 4 mil length orifice in a stainless steel spinneret through a 3.5 inch air gap into the container of isopropanol at 20 °C. The filament which formed as the solution was extruded into 20 the isopropanol at 8.3 fpm and was wound on a bobbin at 11 fpm.

The spun filament was allowed to stand in isopropanol overnight. Then, the filament was drawn while still wet to 2X its length at 150 °C in a tube 25 furnace. The drawn fiber was then allowed to dry in room air.

Physical testing of samples of the dry fiber showed them to be 16.7 denier, with tenacities of 1.22 gpd, elongations of 103.3%, and initial moduli of 40.1 gpd. 30 These figures indicate that the tenacity and modulus of the spun DP-1 spider silk variant fiber compares favorably with those of commercial textile fibers and is therefore considered to be a useful fiber.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:

- (A) NAME: E. I. DU PONT DE NEMOURS AND COMPANY
- (B) STREET: 1007 MARKET STREET
- (C) CITY: WILMINGTON
- (D) STATE: DELAWARE
- (E) COUNTRY: UNITED STATES OF AMERICA
- (F) POSTAL CODE (ZIP): 19898
- (G) TELEPHONE: 302-992-4929
- (H) TELEFAX: 302-773-0164
- (I) TELEX: 6717325

(ii) TITLE OF INVENTION: NOVEL RECOMBINANTLY PRODUCED SPIDER SILK ANALOGS

(iii) NUMBER OF SEQUENCES: 107

(iv) COMPUTER READABLE FORM:

- (A) MEDIUM TYPE: FLOPPY DISK
- (B) COMPUTER: MACINTOSH
- (C) OPERATING SYSTEM: MACINTOSH 6.0
- (D) SOFTWARE: MICROSOFT WORD 4.0

(v) CURRENT APPLICATION DATA:

- (A) APPLICATION NUMBER:

(vi) PRIOR APPLICATION DATA:

- (A) APPLICATION NUMBER: 08/077,600
- (B) FILING DATE: JUNE 15, 1993

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 34 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Xaa Gln Gly Ala Gly Arg
1 5 10 15

Gly Gly Leu Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Ala
20 25 30

Gly Gly

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 15 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Gly Gly
1 5 10 15

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 5 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

Gly Pro Gly Gly Tyr
1 5

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 5 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

Gly Pro Gly Gln Gln
1 5

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 14 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

ACGACCTCAT CTAT 14

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 14 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CTGCCTCTGT CATC 14

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 14 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

AATAGGCGTA TCAC

14

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 19 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

Gly Arg Gly Ala Gly Gln Ser Gly Leu Gly Gly Tyr Gly Gly Gln Gly
1 5 10 15

Ala Gly Cys

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 19 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

Ser Pro Gly Gln Gln Gly Pro Gly Tyr Gly Gly Pro Gly Gln Gln Gly
1 5 10 15

Pro Gly Cys

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 10 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

Gly Ser His His His His His Ser Arg
1 5 10

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GATCCCATCA CCATCACCAT CACTCTA

27

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GATCTAGAGT GATGGTGATG GTGATGG

27

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 8 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: unknown
(D) TOPOLOGY: unknown

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

Gly Ser His His His His His His
1 5

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GATCCCATCA CCATCACCAT CACTAAA

27

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 27 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GATCTTTAGT GATGGTGATG GTGATGG

27

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 12 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GATCAGATAT CG

12

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 12 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GATCCGATAT CT

12

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 47 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly
1 5 10 15

Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro
20 25 30

Ser Gly Pro Gly Ser Ala
35 40 45

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 651 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly
1 5 10 15

Gly Tyr Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Gly Tyr Gly
20 25 30

Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala
35 40 45

Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser
50 55 60

Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala
65 70 75 80

Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly
85 90 95

Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala
100 105 110

Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Asn
115 120 125

Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Ala Ala Ala Ala Gly
130 135 140

Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly
145 150 155 160

Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala
165 170 175

Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Gly Gln Gly Ala
180 185 190

Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly
195 200 205

Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly
210 215 220

Gly Ala Gly Gln Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Ala
225 230 235 240

Gly Ala Ser Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly
245 250 255

Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Glu Gly Ala Gly Ala
260 265 270

Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Leu
275 280 285

Gly Gly Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
290 295 300

Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala
305 310 315 320

Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln
325 330 335

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly
340 345 350

Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly
355 360 365

Gln Gly Ala Gly Ala Val Ala Ala Ala Ala Gly Gly Ala Gly Gln
370 375 380

Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln
385 390 395 400

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Arg Gly
405 410 415

Tyr Gly Gly Leu Gly Asn Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly
420 425 430

Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln
435 440 445

Gly Gly Tyr Gly Gly Leu Gly Asn Gln Gly Ala Gly Arg Gly Gly Gln
450 455 460

Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly
465 470 475 480

Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala
485 490 495

Ala Ala Ala Ala Val Gly Ala Gly Gln Glu Gly Ile Arg Gly Gln Gly
500 505 510

Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ser Gly Arg
515 520 525

Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly
530 535 540

Gly Ala Gly Gln Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Ala
545 550 555 560

Gly Ala Ala Ala Ala Ala Gly Gly Val Arg Gln Gly Gly Tyr Gly
565 570 575

Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala
580 585 590

Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu
595 600 605

Gly Gly Gln Gly Val Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly
610 615 620

Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Val Gly
625 630 635 640

Ser Gly Ala Ser Ala Ala Ser Ala Ala Ala
645 650

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 101 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: unknown
(D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala
1 5 10 15

Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala
20 25 30

Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala
35 40 45

Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly
50 55 60

Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly
 65 70 75 80
 Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly
 85 90 95
 Gly Leu Gly Ser Gln
 100

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 606 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala
 1 5 10 15
 Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala
 20 25 30
 Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala
 35 40 45
 Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly
 50 55 60
 Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly
 65 70 75 80
 Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly
 85 90 95
 Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala
 100 105 110
 Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu
 115 120 125
 Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly
 130 135 140
 Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly
 145 150 155 160
 Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly
 165 170 175
 Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly
 180 185 190

Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly
195 200 205

Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly
210 215 220

Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly
225 230 235 240

Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly
245 250 255

Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala
260 265 270

Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
275 280 285

Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly
290 295 300

Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly
305 310 315 320

Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly
325 330 335

Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala
340 345 350

Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln
355 360 365

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly
370 375 380

Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly
385 390 395 400

Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala
405 410 415

Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
420 425 430

Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala
435 440 445

Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser
450 455 460

Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly
465 470 475 480

Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln
485 490 495

Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln
500 505 510

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly
515 520 525

Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly
530 535 540

Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln
545 550 555 560

Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala
565 570 575

Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser
580 585 590

Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
595 600 605

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 101 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly
1 5 10 15

Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
20 25 30

Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln
35 40 45

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly
50 55 60

Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala
65 70 75 80

Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly
85 90 95

Gly Leu Gly Ser Gln
100

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 606 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: unknown
(D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly
1 5 10 15

Arg Gly Gly Leu Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
20 25 30

Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln
35 40 45

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly
50 55 60

Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala
65 70 75 80

Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly
85 90 95

Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
100 105 110

Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala
115 120 125

Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser
130 135 140

Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly
145 150 155 160

Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg
165 170 175

Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly
180 185 190

Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly
195 200 205

Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly
210 215 220

Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln
225 230 235 240

Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala
245 250 255

Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser
260 265 270

Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala
275 280 285

Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly
290 295 300

Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg
305 310 315 320

Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
325 330 335

Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly
340 345 350

Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr
355 360 365

Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly
370 375 380

Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly
385 390 395 400

Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser
405 410 415

Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala
420 425 430

Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln
435 440 445

Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala
450 455 460

Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly
465 470 475 480

Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln
485 490 495

Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr
500 505 510

Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln
515 520 525

Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly
530 535 540

90

Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala
545 550 555 560

Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
565 570 575

Gly Ala Gly Arg Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
580 585 590

Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
595 600 605

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 93 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

GGGCCGGACG TGGTGGCCTT GGTGGTCAGG GTGCTGGCGC GGCAGCCGCT GCAGCAGCTG 60
GTGGTGCTGG TCAGGGCGGT CTTGGCTCAC AAG 93

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 93 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

GTGAGCCAAG ACCGCCCTGA CCAGCACCAAC CAGCTGCCGC AGCGGCTGCC GCGCCAGCAC 60
CCTGACCAACC AAGGCCACCA CGTCCGGCCC CTT 93

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 31 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala
1 5 10 15

Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln
20 25 30

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 81 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

GGGCCGGTCA AGGCGCTGGT GCAGCAGCAG CTGCCGCTGG CGGTGCAGGC CAAGGTGGAT 60
ATGGTGGCTT AGGGTCACAA G 81

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 81 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

GTGACCCCTAA GCCACCATAT CCACCTTGGC CTGCACCGCC AGCGGCAGCT GCTGCTGCAC 60
CAGCGCCCTTG ACCGGCCCCCT T 81

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: unknown
(D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala
1 5 10 15

Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
 20 . 25

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 90 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

GGGCCGGTCG AGGTGGACAA GGTGCAGGTG CAGCCGCTGC TGCTGCAGGC GGCGCAGGTC 60
 AAGGTGGGTA TGGGGGTTTA GGTCACAAAG 90

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 90 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

GTGAAACCTAA ACCCCCATAAC CCACCTTGAC CTGCGCCGCC CGCAGCAGCA GCGGCTGCAC 60
 CTGCACCTTG TCCACCTCGA CCGGCCCCTT 90

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 30 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS: unknown
 (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
 1 5 10 15

Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
 20 . 25 30

(2) INFORMATION FOR SEQ ID NO:33:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 39 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: DNA (genomic)

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

GGGCCGGGCA AGGTGGTTAC GGCGGTCTCG GATCACAAG

39

(2) INFORMATION FOR SEQ ID NO:34:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 39 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: DNA (genomic)

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

GTGATCCGAG ACCGCCGTAA CCACCTTGCC CGGCCCCCTT

39

(2) INFORMATION FOR SEQ ID NO:35:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 13 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS: unknown
 - (D) TOPOLOGY: unknown

- (ii) MOLECULE TYPE: peptide

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
1 5 10

(2) INFORMATION FOR SEQ ID NO:36:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 36 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

GATCTGCGGC CCAAGGGGCC CACAAGGTGA GG 32

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 32 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

ACGCCGGGTT CCCCGGGTGT TCCACTCCCT AG 32

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 9 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

Ser Ala Ala Gln Gly Ala His Lys Val
1 5

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 42 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

GGATCCCATC ACCATCACCA TCACTCTAGA TCCGGCTGCT AA 42

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 13 amino acids
- (B) TYPE: amino acid

(C) STRANDEDNESS: unknown
(D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

Gly Ser His His His His His Ser Arg Ser Gly Cys
1 5 10

(2) INFORMATION FOR SEQ ID NO:41:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 66 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

GATCTCCCGG GCCATCCGGC CCAGGTTCTG CGGCAGCGGC AGCAGCGGGC CCAGGGCAGC 60
AGCTGG 66

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 66 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

GATCCCAGCT GCTGCCCTGG GCCCGCTGCT GCCGCTGCCG CAGAACCTGG GCCGGATGGC 60
CCGGGA 66

(2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: unknown
(D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

Ser Pro Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Gly
1 5 10 15

Pro Gly Gln Gln Leu
20

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 72 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

GATCTCCGGG GCCGGGCGGT TACGGTCCGG GTCAGCAAGG CCCAGGTGGC TACGGCCCAG 60
GCCAACAGCT GG 72

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 72 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

GATCCCAGCT GTTGGCCTGG GCCGTAGCCA CCTGGGCCTT GCTGACCCGG ACCGTAACCG 60
CCCGGGCCCGG GA 72

(2) INFORMATION FOR SEQ ID NO:46:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 23 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

Ser Pro Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly
1 5 10 15

Tyr Gly Pro Gly Gln Gln Leu
20

(2) INFORMATION FOR SEQ ID NO:47:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 72 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

GATCTCCCGG GCCATCTGGT CCGGGTAGCG CTGCGGCTGC TGCTGCTGCG GCAGGGTCCAG 60
GCGGGCTACGT AG 72

(2) INFORMATION FOR SEQ ID NO:48:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 72 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

GATCCTACGT AGCCGCCTGG ACCTGCCGCA GCAGCAGCAG CCGCAGCGCT ACCCGGACCA 60
GATGGCCCCGG GA 72

(2) INFORMATION FOR SEQ ID NO:49:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 23 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 49:

Ser Pro Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Ala Ala
1 5 10 15
Ala Gly Pro Gly Gly Tyr Val
20

(2) INFORMATION FOR SEQ ID NO:50:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 57 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

GATCTCCCGG GCCGGGCCAA CAAGGTCCGG GCGGCTATGG TCCAGGTCAA CAGCTGG 57

(2) INFORMATION FOR SEQ ID NO:51:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 57 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

GATCCCAGCT GTTGACCTGG ACCATAGCCG CCCGGACCTT GTTGGCCCGG CCCGGGA 57

(2) INFORMATION FOR SEQ ID NO:52:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: unknown
(D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

Ser Pro Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln
1 5 10 15

Gln Leu

(2) INFORMATION FOR SEQ ID NO:53:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 75 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:

GATCTCCCGG GCCGAGCGGT CCAGGTTCCG CAGCAGCAGC GGCTGCGGCG GCAGCGGGTC 60
CAGGTGGTTA CGTAG 75

(2) INFORMATION FOR SEQ ID NO:54:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 75 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

GATCCTACGT AACCACCTGG ACCCGCTGCC GCCGCAGCCG CTGCTGCTGC GGAACCTGGA 60
CCGCTCGGCC CGGGGA 75

(2) INFORMATION FOR SEQ ID NO:55:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:

Ser Pro Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Ala
1 5 10 15

Ala Ala Gly Pro Gly Gly Tyr Val
20

(2) INFORMATION FOR SEQ ID NO:56:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 87 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

100

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

GATCTCCCGG GCCAGGCCAG CAGGGTCCGG GTGGCTATGG CCCAGGCCAG CAAGGTCCGG	60
GTGGTTACGG TCCAGGTCAG CAGCTGG	87

(2) INFORMATION FOR SEQ ID NO:57:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 87 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

GATCCCAGCT GCTGACCTGG ACCGTAACCA CCCGGACCTT GCTGGCCTGG GCCATAGCCA	60
CCCGGACCTT GCTGGCCTGG CCCGGGA	87

(2) INFORMATION FOR SEQ ID NO:58:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 28 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

Ser Pro Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln			
1	5	10	15
Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Leu			
20	25		

(2) INFORMATION FOR SEQ ID NO:59:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 493 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:

Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly			
1	5	10	15

Pro Gly Gln Gln Gly Pro Gly Arg Tyr Gly Pro Gly Gln Gln Gly Pro
20 25 30

Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Gly Ser Gly Gln Gln
35 40 45

Gly Pro Gly Gly Tyr Gly Pro Arg Gln Gln Gly Pro Gly Gly Tyr Gly
50 55 60

Gln Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ser
65 70 75 80

Ala Ala Ala Ser Ala Glu Ser Gly Gly Pro Gly Gly Tyr Gly Pro Gly
85 90 95

Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly
100 105 110

Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala
115 120 125

Ala Ala Ala Ala Ala Ser Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr
130 135 140

Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly
145 150 155 160

Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Ala Ser Gly
165 170 175

Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro
180 185 190

Gly Gly Tyr Gly Pro Gly Gln Gln Gly Thr Ser Gly Pro Gly Ser Ala
195 200 205

Ala Ala Ala Ala Ala Ala Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr
210 215 220

Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala
225 230 235 240

Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly
245 250 255

Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser
260 265 270

Ala Ala Ala Ala Ala Ala Gly Pro Gly Gln Gln Gly Leu Gly Gly
275 280 285

Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln
290 295 300

Gly Pro Gly Gly Tyr Gly Pro Gly Ser Ala Ser Ala Ala Ala Ala Ala
305 310 315 320

102

Ala Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln
325 330 335

Gly Pro Ser Gly Pro Gly Ser Ala Ser Ala Ala Ala Ala Ala Ala
340 345 350

Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr
355 360 365

Ala Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ser Ala Ala
370 375 380

Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln
385 390 395 400

Gly Pro Gly Gly Tyr Ala Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly
405 410 415

Ser Ala Ala Ala Ala Ala Ala Ser Ala Gly Pro Gly Gly Tyr Gly
420 425 430

Pro Ala Gln Gln Gly Pro Ser Gly Pro Gly Ile Ala Ala Ser Ala Ala
435 440 445

Ser Ala Gly Pro Gly Gly Tyr Gly Pro Ala Gln Gln Gly Pro Ala Gly
450 455 460

Tyr Gly Pro Gly Ser Ala Val Ala Ala Ser Ala Gly Ala Gly Ser Ala
465 470 475 480

Gly Tyr Gly Pro Gly Ser Gln Ala Ser Ala Ala Ala Ser
485 490

(2) INFORMATION FOR SEQ ID NO:60:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 119 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:

Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Gly Pro Gly
1 5 10 15

Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly
20 25 30

Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala
35 40 45

Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly
50 55 60

103

Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser
 65 70 75 80

Ala Ala Ala Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro
 85 90 95

Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly
 100 105 110

Gly Tyr Gly Pro Gly Gln Gln
 115

(2) INFORMATION FOR SEQ ID NO:61:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 714 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Gly Pro Gly
 1 5 10 15

Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly
 20 25 30

Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala
 35 40 45

Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly
 50 55 60

Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser
 65 70 75 80

Ala Ala Ala Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro
 85 90 95

Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly
 100 105 110

Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala
 115 120 125

Ala Ala Ala Ala Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro
 130 135 140

Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser
 145 150 155 160

Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Gly Pro Gly Gly
 165 170 175

104

Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln
180 185 190

Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Ala Ala
195 200 205

Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly
210 215 220

Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro
225 230 235 240

Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Gly Pro Gly Gln Gln
245 250 255

Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly
260 265 270

Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala
275 280 285

Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly
290 295 300

Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala
305 310 315 320

Ala Ala Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln
325 330 335

Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr
340 345 350

Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala
355 360 365

Ala Ala Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln
370 375 380

Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro
385 390 395 400

Gly Ser Ala Ala Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly
405 410 415

Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro
420 425 430

Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala Ala Ala Gly Pro
435 440 445

Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly
450 455 460

Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly
465 470 475 480

105

Pro Gly Ser Ala Ala Ala Ala Ala Gly Pro Gly Gln Gln Gly Pro
485 490 495

Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly
500 505 510

Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala Ala
515 520 525

Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr
530 535 540

Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala
545 550 555 560

Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly
565 570 575

Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro
580 585 590

Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser Ala Ala Ala Ala Ala
595 600 605

Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly
610 615 620

Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly Pro Gly Ser
625 630 635 640

Ala Ala Ala Ala Ala Ala Ala Gly Pro Gly Gly Tyr Gly Pro Gly
645 650 655

Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln Gly Pro Ser Gly
660 665 670

Pro Gly Ser Ala Ala Ala Ala Ala Ala Ala Ala Ala Gly Pro Gly Gly
675 680 685

Tyr Gly Pro Gly Gln Gln Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln
690 695 700

Gly Pro Gly Gly Tyr Gly Pro Gly Gln Gln
705 710

(2) INFORMATION FOR SEQ ID NO:62:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 101 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

106

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly
1 5 10 15

Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala
20 25 30

Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala
35 40 45

Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln
50 55 60

Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln
65 70 75 80

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly
85 90 95

Tyr Gly Gly Leu Gly
100

(2) INFORMATION FOR SEQ ID NO:63:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 604 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:

Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly
1 5 10 15

Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala
20 25 30

Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala
35 40 45

Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln
50 55 60

Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln
65 70 75 80

Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly
85 90 95

Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly
100 105 110

107

Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala
115 120 125

Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu
130 135 140

Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
145 150 155 160

Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala
165 170 175

Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly
180 185 190

Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln
195 200 205

Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu
210 215 220

Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala
225 230 235 240

Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala
245 250 255

Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu
260 265 270

Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala
275 280 285

Ala Ala Ala Gly Gly Ala Gly Gly Tyr Gly Gly Leu Gly Ser Gly
290 295 300

Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg
305 310 315 320

Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
325 330 335

Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly
340 345 350

Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr
355 360 365

Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly
370 375 380

Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly
385 390 395 400

Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser
405 410 415

108

Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala
420 425 430

Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Ser Gln
435 440 445

Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala
450 455 460

Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly
465 470 475 480

Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln
485 490 495

Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Gly Tyr
500 505 510

Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln
515 520 525

Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly
530 535 540

Gly Leu Gly Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala
545 550 555 560

Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln
565 570 575

Gly Ala Gly Arg Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala
580 585 590

Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
595 600

(2) INFORMATION FOR SEQ ID NO:64:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:

GATCTCAGGG TGCTGGCCAG GGTGGCTATG GTGGCCTGG 39

(2) INFORMATION FOR SEQ ID NO:65:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 base pairs
- (B) TYPE: nucleic acid

109

- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:

GATCCCAGGC CACCATAGCC ACCCTGGCCA GCACCCTGA

39

(2) INFORMATION FOR SEQ ID NO:66:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 13 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS: unknown
 - (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:

Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
1 5 10

(2) INFORMATION FOR SEQ ID NO:67:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 93 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:

GATCTCAAGG CGCTGGTCGC GGTGGCCTGG GTGGCCAGGG TGCAGGTGCT GCTGCTGCTG 60

CGGCTGCTGG TGGTGCAGGT CAGGGTGGTC TGG 93

(2) INFORMATION FOR SEQ ID NO:68:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 93 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:

GATCCCAGAC CACCTGTGACC TGCACCA GCAGCCGCAG CAGCAGCAGC ACCTGCACCC 60

110

TGGCCACCCA GGCCACCGCG ACCAGGCCT TGA

93

(2) INFORMATION FOR SEQ ID NO:69:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 31 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:69:

Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala
1 5 10 15

Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly
20 25 30

(2) INFORMATION FOR SEQ ID NO:70:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 81 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:70:

GATCTCAGGG CGCAGGTCAA GGTGCTGGTG CAGCTGCGGC GGCAGCTGGT GGCGCGGGTC 60

AAGGTGGCTA CGGCGGTTA G 81

(2) INFORMATION FOR SEQ ID NO:71:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 81 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:71:

GATCCTAAC CGCCGTAGCC ACCTTGACCC GCGCCACCAG CTGCGCCGC AGCTGCACCA 60

GCACCTTGAC CTGCGCCCTG A 81

(2) INFORMATION FOR SEQ ID NO:72:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 28 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:72:

Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly
1 5 10 15

Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
20 25

(2) INFORMATION FOR SEQ ID NO:73:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 90 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:73:

GATCTCAAGG TGCAGGTTCGC GGTGGTCAGG GCGCTGGTGC AGCAGCGGCA GCAGCAGGTG 60
GCGCTGGCCA AGGTGGTTAC GGTGGTCTTG 90

(2) INFORMATION FOR SEQ ID NO:74:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 90 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:74:

GATCCAAGAC CACCGTAACC ACCTTGGCCA GCGCCACCTG CTGCTGCCGC TGCTGCACCA 60
GCGCCCTGAC CACCGCGACC CGCACCTTGA 90

(2) INFORMATION FOR SEQ ID NO:75:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 amino acids

112

- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:75:

Ser Gln Gly Ala Gly Arg Gly Gln Gly Ala Gly Ala Ala Ala Ala
1 5 10 15
Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
20 25 30

(2) INFORMATION FOR SEQ ID NO:76:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 18 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:76:

AATTCAAGATC TAAGCTTG 18

(2) INFORMATION FOR SEQ ID NO:77:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 18 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:77:

GATCCAAGCT TAGATCTG 18

(2) INFORMATION FOR SEQ ID NO:78:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 4909 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: circular

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:78:

GAATTCCGGG GGATTATGCG TTAAGCATAA AGTGTAAAGC CTGGGGTGCC TAATGAGTGA 60
GCTAACTCAC ATTAATTGCG TTGCGCTCAC TGCCCCTTT CCAGTCGGGA AACCTGTCGT 120
GCCAGCTGCA TTAATGAATC GGCCAACGCG CGGGGAGAGG CGGTTTGCCT ATTGGGCGCC 180
AGGGTGGTTT TTCTTTTCAC CAGTGAGACG GGCAACAGCT GATTGCCCTT CACCGCCTGG 240
CCCTGAGAGA GTTGCAGCAA GCGGTCACG CTGGTTTGCC CCAGCAGGCG AAAATCCTGT 300
TTGATGGTGG TTGACGGCGG GATATAACAT GAGCTGTCTT CGGTATCGTC GTATCCCACT 360
ACCGAGATAT CCGCACCAAC GCGCAGCCCG GACTCGGTAA TGGCGCGCAT TGGCCCGAGC 420
GCCATCTGAT CGTTGGCAAC CAGCATCGCA GTGGGAACGA TGCCCTCATT CAGCATTG 480
ATGGTTTGTGTT GAAAACCGGA CATGGCACTC CAGTCGCCTT CCCGTTCCGC TATCGGCTGA 540
ATTTGATTGC GAGTGAGATA TTTATGCCAG CCAGCCAGAC GCAGACGCGC CGAGACAGAA 600
CTTAATGGC CCGCTAACAG CGCGATTTGC TGGTGACCCA ATGCGACCAAG ATGCTCCACG 660
CCCAGTCGCG TACCGTCTTC ATGGGAGAAA ATAATACTGT TGATGGGTGT CTGGTCAGAG 720
ACATCAAGAA ATAACGCCGG AACATTAGTG CAGGCAGCTT CCACAGCAAT GGCATCCTGG 780
TCATCCAGCG GATAGTTAAT GATCAGCCCA CTGACGCGTT GCGCGAGAAG ATTGTGCACC 840
GCCGCTTAC AGGCTTCGAC GCCGCTTCGT TCTACCATCG ACACCACCAC GCTGGCACCC 900
AGTTGATCGG CGCGAGATT AATGCCGCG ACAATTGCG ACGGCCGCG CAGGGCCAGA 960
CTGGAGGTGG CAACGCCAAT CAGCAACGAC TGTTTGCCCG CCAGTTGTTG TGCCACGCGG 1020
TTGGGAATGT AATTCACTGCT CGCCATCGCC GCTTCCACTT TTTCCCGCGT TTTCGCAGAA 1080
ACGTGGCTGG CCTGGTTCAC CACGCCGGAA ACGGTCTGAT AAGAGACACC GGCATACTCT 1140
GCGACATCGT ATAACGTTAC TGGTTTCACA TTCACCACCC TGAATTGACT CTCTTCCGGG 1200
CGCTATCATG CCATACCGCG AAAGGTTTG CGCCATTGCA TGGTGTCAAC CTTGCAGAGC 1260
TGCCTCTTA TTATTATCCG CCGGGAGAAA ATATTCCGTG GATCTAACGG GATGCGTTAT 1320
GTTGAAGTGA GACCGGTCGA CGCATGCCAG GACAACCTCT GGTCCGGTAA CGTGCTGAGC 1380
CCGGCCAAGC TTACTCCCCA TCCCCCTGTT GACAATTAAAT CATCGGCTCG TATAATGTGT 1440
GGAATTGTGA GCGGATAACA ATTCACACCA GGAAACAGGA TCACTAAGGA GGTTAAATA 1500
TGGCTACTGT TATAGATCCG TCTGTCGCGA CGGCCGTTTC GTCGAATGGC TCGGTTGCCA 1560
ATATCAATGC GATCAAGTCG GGCCTCTGG AGTCCGGCTT TACGCAGTCA GACGTTGCCT 1620
ATTGGGCCTA TAACGGCACC GGCCTTATG ATGGCAAGGG CAAGGTGGAA GATTTGCGCC 1680

TTCTGGCGAC GCTTTACCCG GAAACGATCC ATATCGTTGC GCGTAAGGAT GCAAACATCA 1740
AATCGGTCGC AGACCTGAAA GGCAAGCGCG TTTCGCTGGA TGAGCCGGGT TCTGGCACCA 1800
TCGTCGATGC GCGTATCGTT CTTGAAGCCT ACGGCCTCAC GGAAGACGAT ATCAAGGCTG 1860
AACACCTGAA GCCGGGACCG GCAGGGAGA GGCTGAAAGA TGGTGCGCTG GACGCCTATT 1920
TCTTTGTGGG CGGCTATCCG ACGGGCGCAA TCTCGGAACG GGCCATCTCG AACGGTATTT 1980
CGCTCGTTCC GATCTCCGGG CCGGAAGCGG ACAAGATTCT GGAGAAATAT TCCTTCTTCT 2040
CGAAGGATGT GGTTCTGCC GGAGCCTATA AGGACGTGGC GGAAACACCG ACCCTTGCCG 2100
TTGCCGCACA GTGGGTGACG AGCGCCAAGC AGCCGGACGA CCTCATCTAT AACATCACCA 2160
AGGCTGGTTC TCCGAAACCG GGTGCTGGTA GATCTAAGCT TCCCGGGGAT CCTAGCTAGC 2220
TAGCCATGGC ATCACAGTAT CGTGATGACA GAGGCAGGGA GTGGGACAAA ATTGAAATCA 2280
ATAATGATT TTATTTGAC TGATAGTGAC CTGTTCGTTG CAACAAATTG ATAAGCAATG 2340
CTTTTTATA ATGCCAACTT AGTATAAAAA AGCTGAACGA GAAACGTAAA ATGATATAAA 2400
TATCAATATA TTAAATTAGA TTTGCATAA AAAACAGACT ACATAATACT GTAAAACACA 2460
ACATATGCAG TCACTATGAA TCAACTACTT AGATGGTATT AGTGACCTGT AACAGAGCAT 2520
TAGCGCAAGG TGATTTTGT CTTCTTGGC TAATTTTTG TCATCAAACC TGTGCACTC 2580
CAGAGAAGCA CAAAGCCTCG CAATCCAGTG CAAAGCTCTG CCTCGCGCGT TTGGTGATG 2640
ACGGTGAAAA CCTCTGACAC ATGCAGCTCC CGGAGACGGT CACAGCTTGT CTGTAAGCGG 2700
ATGCCGGGAG CAGACAAGCC CGTCAGGGCG CGTCAGCGGG TGTTGGCGGG TGTCGGGGCG 2760
CAGCCATGAC CCAGTCACGT AGCGATAGCG GAGTGTATAC TGGCTTAACG ATGCGGCATC 2820
AGAGCAGATT GTACTGAGAG TGCACCATAT GCGGTGTGAA ATACCGCACA GATGCGTAAG 2880
GAGAAAATAC CGCATCAGGC GCTCTTCCGC TTCCCTCGCTC ACTGACTCGC TGCCTCGGT 2940
CGTTCGGCTG CGCGAGCGG TATCAGCTCA CTCAAAGGCG GTAATACGGT TATCCACAGA 3000
ATCAGGGGAT AACGCAGGAA AGAACATGTG AGCAAAAGGC CAGCAAAAGG CCAGGAACCG 3060
TAAAAAGGCC GCGTTGCTGG CGTTTTCCA TAGGCTCCGC CCCCCCTGACG AGCATCACAA 3120
AAATCGACGC TCAAGTCAGA GGTGGCGAAA CCCGACAGGA CTATAAAAGAT ACCAGGCGTT 3180
TCCCCCTGGA AGCTCCCTCG TGCCTCTCC TGTTCCGACC CTGCCGCTTA CCGGATACCT 3240
GTCCGCCTT CTCCCTTCGG GAAGCGTGGC GCTTTCTCAT AGCTCACGCT GTAGGTATCT 3300
CAGTTCGGTG TAGGTCGTTG GCTCCAAGCT GGGCTGTGTG CACGAACCCC CCGTTCAGCC 3360
CGACCGCTGC GCCTTATCCG GTAACTATCG TCTTGAGTCC AACCCGGTAA GACACGACTT 3420

ATCGCCACTG GCAGCAGCCA CTGGTAACAG GATTAGCAGA GCGAGGTATG TAGGCGGTGC 3480
TACAGAGTTC TTGAAGTGGT GGCTAACTA CGGCTACACT AGAAGGACAG TATTTGGTAT 3540
CTGCGCTCTG CTGAAGCCAG TTACCTCGG AAAAAGAGTT GGTAGCTCTT GATCCGGCAA 3600
ACAAACCACC GCTGGTAGCG GTGGTTTTT TGTGCAAG CAGCAGATTA CGCGCAGAAA 3660
AAAAGGATCT CAAGAAGATC CTTTGATCTT TTCTACGGGG TCTGACGCTC AGTGGAACGA 3720
AAACTCACGT TAAGGGATT TGGTCATGAG ATTATCAAAA AGGATCTTCA CCTAGATCCT 3780
TTTAAATTAA AAATGAAGTT TTAAATCAAT CTAAGTATA TATGAGTAAA CTTGGTCTGA 3840
CAGTTACCAA TGCTTAATCA GTGAGGCACC TATCTCAGCG ATCTGTCTAT TTGTTCATC 3900
CATAGTTGCC TGACTCCCCG TCGTGTAGAT AACTACGATA CGGGAGGGCT TACCATCTGG 3960
CCCCAGTGCT GCAATGATAC CGCGAGACCC ACGCTCACCG GCTCCAGATT TATCAGCAAT 4020
AAACCAGCCA GCCGGAAGGG CCGAGCGCAG AAGTGGTCCT GCAACTTTAT CCGCCTCCAT 4080
CCAGTCTATT AATTGTTGCC GGGAGCTAG AGTAAGTAGT TCGCCAGTTA ATAGTTGCG 4140
CAACGTTGTT GCCATTGCTG CAGGCATCGT GGTGTCACGC TCGTCGTTG GTATGGCTTC 4200
ATTCAAGCTCC GGTTCCCAAC GATCAAGGCG AGTTACATGA TCCCCCATGT TGTGCAAAAA 4260
AGCGGTTAGC TCCCTCGGTC CTCCGATCGT TGTCAAGGT AAGTTGGCCG CAGTGTATC 4320
ACTCATGGTT ATGGCAGCAC TGCATAATT TCTTACTGTC ATGCCATCCG TAAGATGCTT 4380
TTCTGTGACT GGTGAGTACT CAACCAAGTC ATTCTGAGAA TAGTGTATGC GGCGACCGAG 4440
TTGCTCTTGC CCGCGTCAA CACGGATAA TACCGCGCCA CATAGCAGAA CTTAAAAGT 4500
GCTCATCATT GGAAAACGTT CTTGGGGCG AAAACTCTCA AGGATCTTAC CGCTGTTGAG 4560
ATCCAGTTCG ATGTAACCCA CTCGTGCACC CAACTGATCT TCAGCATCTT TTACTTTCAC 4620
CAGCGTTCT GGGTGAGCAA AAACAGGAAG GCAAAATGCC GCAAAAAGG GAATAAGGGC 4680
GACACGGAAA TGTTGAATAC TCATACTCTT CCTTTTCAA TATTATTGAA GCATTTATCA 4740
GGGTTATTGT CTCATGAGCG GATACATATT TGAATGTATT TAGAAAAATA AACAAATAGG 4800
GGTTCCGCGC ACATTTCCCC GAAAAGTGC ACCTGACGTC TAAGAAACCA TTATTATCAT 4860
GACATTAACC TATAAAAATA GGCGTATCAC GAGGCCCTT CGTCTCAA 4909

(2) INFORMATION FOR SEQ ID NO:79:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 9144 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: circular

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:79:

AATTCGAGCT CGGTACCCAT CGAATTCTT CAGGAAAAGA ACGATGGCTG TCTTATTAGC 60
GGTTGCAGGC ACATTTATTG TGGTCACACA CGGGAAATGTC GGCAGCCTGT CTATATCCGG 120
TCTGGCTGTT TTTTGGGGCA TCAGCTCGGC ATTTGCGCTG GCGTTTACA CCCTCCAGCC 180
GCATCGGCTT TTGAAGAAAT GGGGCTCCGC CATTATTGTC GGATGGGGCA TGCTGATGCG 240
GAGCCGTTCT CAGCCTGATT CAGCCGCCCTT GGAAGTTGA AGGCCAATGG TCGTTGTCCG 300
CATATGCCGC GATCGTGTTC ATCATCATT TCAGGAACGCT CATCGCTTT TATTGCTATT 360
TGGAAAGCCT GAAATATCTG AGTGCCTCTG AAACCAGCCT CCTCGCCTGT GCAGAGCCGC 420
TGTCAAGCAGC TTTTTAGCG GTGATCTGGC TGCAATGTTCC CTTCGGAATA TCAGAATGGC 480
TGGGTACTTT ACTGATTTTA GCCACCATCG CTTATTATCT ATCAAGAAAA AATAACCTCT 540
CTTTTTTAG AGAGGTTTT CCCTAGGCCT GAAGCACCCCT TTAGTCTCAA TTACCCATAA 600
ATTAAAAGGC CTTTTTCGT TTTACTATCA TTCAAAAGAG GAAAATAGAC CAGTTGTCAA 660
TAGAACATAGA GTCTAATAGA ATGAGGTGCA AAAGTAAATC ACGCAGGATT GTTACTGATA 720
AAGCAGGCAA GACCTAAAAT GTGTTAAGGG CAAAGTGTAT TCTTTGGCGT CATCCCTTAC 780
ATATTTGGG TCTTTTTTC TGTAACAAAC CTGCCATCCA TGAATTGGGG AGGATCGAAA 840
CGGCAGATCG CAAAAACAGT ACATACAGAA GGAGACATGA ACATGAACAT CAAAAAAATT 900
GTAAAACAAG CCACAGTACT GACTTTACG ACTGCACTGC TAGCAGGAGG AGCGACTCAA 960
GCCTTCGCGA AAGAAGATAT CGATCAACGC AATGGTTTA TCCAAAGCCT TAAAGATGAT 1020
CCAAGCCAAA GTGCTAACGT TTTAGGTGAA GCTAAAAAC TTAATGACTC TCAAGCTCCA 1080
AAAGCTGATG CGCAACAAAA TAACTTCAAC AAAGATCAAC AAAGCGCCTT CTATGAAATC 1140
TTGAACATGC CTAACCTAAA CGAAGCGCAA CGTAACGGCT TCATTCAAAG TCTTAAAGAC 1200
GACCCAAGCC AAAGCACTAA CGTTTAGGT GAAGCTAAAA AATTAAACGA ATCTCAAGCA 1260
CCGAAAGCTG ATAACAATT CAACAAAGAA CAACAAAATG CTTCTATGA AATCTTGAAT 1320
ATGCCTAACT TAAACGAAGA ACAACGCAAT GGTTTCATCC AAAGCTTAAA AGATGACCCA 1380
AGCCAAAGTG CTAACCTATT GTCAGAAGCT AAAAAGTTAA ATGAATCTCA AGCACCGAAA 1440
GCGGATAACA AATTCAACAA AGAACAAACAA AATGCTTTCT ATGAAATCTT ACATTTACCT 1500
AACTTAAACG AAGAACAAACG CAATGGTTTC ATCCAAAGCC TAAAAGATGA CCCAAGCCAA 1560

AGCGCTAACCTTTAGCAGA AGCTAAAAAG CTAATGATG CTCAAGCACC AAAAGCTGAC 1620
AACAAATTCA ACAAAAGAAC ACAAATGCT TTCTATGAAA TTTTACATTT ACCTAACTTA 1680
ACTGAAGAAC AACGTAACGG CTTCATCCAA AGCCTTAAAG ACGATCCGGG GAATTCCGG 1740
GGATCCGTCG ACCTGCAGGC ATGCAAGCTT ACTCCCCATC CCCTCCAGTA ATGACCTCAG 1800
AACTCCATCT GGATTGTTG AGAACCGCTCG GTTGCAGCCG GGCGTTTTT ATTGGTGAGA 1860
ATCGCAGCAA CTTGTCGCGC CAATCGAGCC ATGTCGTCGT CAACGACCCC CCATTCAAGA 1920
ACAGCAAGCA GCATTGAGAA CTTTGAATC CAGTCCCTCT TCCACCTGCT GAGGGCAATA 1980
AGGGCTGCAC GCGCACTTT ATCCGCCTCT GCTGCGCTCC GCCACCGTAG TTAAATTTAT 2040
GGTTGGTTAT GAAATGCTGG CAGAGACCA GCGAGACCTG ACCGCAGAAC AGGCAGCAGA 2100
GCGTTGCGC GCAGTCAGCG ATACCCCGT TGATAATCAG AAAAGCCCCA AAAACAGGAA 2160
GATTGTATAA GCAAATATTT AAATTGAAA CGTTAATATT TTGTTAAAAT TCGCGTTAAA 2220
TTTTGTTAA ATCAGCTCAT TTTTAACCA ATAGGCCGAA ATCGGAAAAA TCCCTTATAA 2280
ATCAAAAGAA TAGCCCGAGA TAGGGTTGAG TGGTGTCCA GTTGGAAACA AGAGTCCACT 2340
ATTAAAGAAC GTGGACTCCA ACGTCAAAGG GCGAAAAACC GTCTATCAGG GCGATGGCCC 2400
ACTACGTGAA CCATCACCCCA AATCAAGTTT TTTGGGGTCG AGGTGCCGTA AAGCACTAAA 2460
TCGGAACCCCT AAAGGGAGCC CCCGATTTAG AGCTTGACGG GGAAAGCCGG CGAACGTGGC 2520
GAGAAAGGAA GGGAAAGAAAG CGAAAGGAGC GGGCGCTAGG GCGCGAGCAA GTGTAGCGGT 2580
CACGCGCGCG TAACCACCAAC ACCCGCCGCG CTTAATGCGC CGCTACAGGG CGCGTATCCA 2640
TTTCGCGAA TCCGGAGTGT AAGAAATGAG TCTGAAAGAA AAAACACAAAT CTCTGTTGC 2700
CAACGCATTT GGCTACCCCTG CCACTCACAC CATTCAAGGTG CGTCATATAC TGACTGAAAA 2760
CGCCCGCACC GTTGAAGCTG CCAGCGCGCT GGAGCAAGGC GACCTGAAAC GTATGGCGA 2820
GTTGATGGCG GAGTCTCATG CCTCTATGCG CGATGATTTC GAAATCACCG TGCCGAAAT 2880
TGACACTCTG GTAGAAATCG TCAAAGCTGT GATTGGCGAC AAAGGTGGCG TACGCATGAC 2940
CGGGCGCGGA TTTGGCGGCT GTATCGTCGC CGGTATCCCG GAAGAGCTGG TGCCCTGCCGC 3000
ACAGCAAGCT GTCGCTGAAC AATATGAAGC AAAAACAGGT ATTAAAGAGA CTTTTTACGT 3060
TTGTAAACCA TCACAAGGAG CAGGACAGTG CTGAACGAAA CTCCCGCACT GGCACCCGAT 3120
GGCAGCCGTA CCGACTGTTC TGCCCTCGCGC GTTTCGGTGA TGACGGTGAA AACCTCTGAC 3180
ACATGCAGCT CCCGGAGACG GTCACAGCTT GTCTGTAAGC GGATGCCGGG AGCAGACAAG 3240
CCCGTCAGGG CGCGTCAGCG GGTGTTGGCG GGTGTCGGGG CGCAGCCATG ACCCAGTCAC 3300

GTAGCGATAG CGGAGTGTAT ACTGGCTTAA CTATGCGGCA TCAGAGCAGA TTGTACTGAG 3360
AGTGCACCAT ATGCGGTGTG AAATACCGCA CAGATGCGTA AGGAGAAAAT ACCGCATCAG 3420
GCGCTCTTCC GCTTCCTCGC TCACTGACTC GCTGCGCTCG GTCGTTGGC TGCGCGAGC 3480
GGTATCAGCT CACTCAAAGG CGGTAATAACG GTTATCCACA GAATCAGGGG ATAACGCAGG 3540
AAAGAACATG TGAGCAAAAG GCCAGCAAA GGCCAGGAAC CGTAAAAAAGG CCGCGTTGCT 3600
GGCGTTTTC CATAGGCTCC GCCCCCCCTGA CGAGCATCAC AAAAATCGAC GCTCAAGTCA 3660
GAGGTGGCGA AACCCGACAG GACTATAAG ATACCAGGCG TTTCCCCCTG GAAGCTCCCT 3720
CGTGCCTCT CCTGTTCCGA CCCTGCCGCT TACCGGATAC CTGTCGCGCT TTCTCCCTTC 3780
GGGAAGCGTG GCGCTTTCTC ATAGCTCACG CTGTAGGTAT CTCAGTTCGG TGTAGGTCGT 3840
TCGCTCCAAG CTGGGCTGTG TGACACGAAACC CCCCGTTCAAG CCCGACCGCT GCGCCTTATC 3900
CGGTAACTAT CGTCTTGAGT CCAACCCGGT AAGACACGAC TTATGCCAC TGGCAGCAGC 3960
CACTGGTAAC AGGATTAGCA GAGCGAGGTA TGTAGGCGGT GCTACAGAGT TCTTGAAGTG 4020
GTGGCCTAAC TACGGCTACA CTAGAAGGGAC AGTATTTGGT ATCTGCCCTC TGCTGAAGCC 4080
AGTTACCTTC GGAAAAAGAG TTGGTAGCTC TTGATCCGGC AAACAAACCA CCGCTGGTAG 4140
CGGTGGTTTT TTTGTTTGCA AGCAGCAGAT TACGCCAGA AAAAAAGGAT CTCAAGAAGA 4200
TCCTTGATC TTTTCTACGG GGTCTGACGC TCAGTGGAAC GAAAACTCAC GTTAAGGGAT 4260
TTTGGTCATG AGATTATCAA AAAGGATCTT CACCTAGATC CTTTTAAATT AAAAAATGAAG 4320
TTTTAAATCA ATCTAAAGTA TATATGAGTA AACTTGGTCT GACAGTTACC AATGCTTAAT 4380
CAGTGAGGCA CCTATCTCAG CGATCTGTCT ATTCGTTCA TCCATAGTTG CCTGACTCCC 4440
CGTCGTGTAG ATAACATACGA TACGGGAGGG CTTACCATCT GGCCCCAGTG CTGCAATGAT 4500
ACCGCGAGAC CCACGCTCAC CGGCTCCAGA TTTATCAGCA ATAAACCAAGC CAGCCGGAAG 4560
GGCCGAGCGC AGAAGTGGTC CTGCAACTTT ATCCGCTCC ATCCAGTCTA TTAATTGTTG 4620
CCGGGAAGCT AGAGTAAGTA GTTCGCCAGT TAATAGTTG CGAACACGTTG TTGCCATTGC 4680
TACAGGCATC GTGGTGTAC GCTCGTCGTT TGGTATGGCT TCATTCAAGCT CCGGTTCCCA 4740
ACGATCAAGG CGAGTTACAT GATCCCCAT GTTGTGCAA AAAGCGGTTA GCTCCTTCGG 4800
TCCTCCGATC GTTGTCAAGAA GTAAAGTTGGC CGCAGTGTAA TCACTCATGG TTATGGCAGC 4860
ACTGCATAAT TCTCTTACTG TCATGCCATC CGTAAGATGC TTTCTGTGA CTGGTGAGTA 4920
CTCAACCAAG TCATTCTGAG AATAGTGTAT GCGGCGACCG AGTTGCTCTT GCCCGGCGTC 4980
AACACGGGAT AATACCGCGC CACATAGCAG AACTTTAAAA GTGCTCATCA TTGGAAAACG 5040

TTCTTCGGGG CGAAAACTCT CAAGGATCTT ACCGCTGTTG AGATCCAGTT CGATGTAACC 5100
CACTCGTGCA CCCAACTGAT CTTCAGCATC TTTTACTTTC ACCAGCGTTT CTGGGTGAGC 5160
AAAAACAGGA AGGCAAAATG CCGCAAAAAA GGGAAATAAGG GCGACACGGA AATGTTGAAT 5220
ACTCATACTC TTCCTTTTC AATATTATTG AAGCATTAT CAGGGTTATT GTCTCATGAG 5280
CGGATACATA TTTGAATGTA TTTAGAAAAA TAAACAAATA GGGGTTCCGC GCACATTTCC 5340
CCGAAAAGTG CCACCTGACG TCTAAGAAC CATTATTATC ATGACATTA CCTATAAAAA 5400
TAGGCGTATC ACGAGGCCCT TTCGTCTCA AGCCCGAGGT AACAAAAAAA CAACAGCATA 5460
AATAACCCCG CTCTTACACA TTCCAGCCCT GAAAAAGGGC ATCAAATTAA ACCACACCTA 5520
TGGTGTATGC ATTTATTTC ATACATTCAA TCAATTGTTA TCTAAGGAAA TACTTACATA 5580
TGGTTCGTGC AAACAAACGC AACGAGGCTC TACGAATCGA TGCATGCAGC TGATTTCACT 5640
TTTTGCATTC TACAAACTGC ATAACTCATA TGAAATCGC TCCTTTTAG GTGGCACAAA 5700
TGTGAGGCAT TTTCGCTCTT TCCGGCAACC ACTTCCAAGT AAAGTATAAC ACACATAC 5760
TTATATTCAAAAGTGTGTG CTCTCGGAGG CTGTCGGCAG TGCCGACCAA AACCATAAAA 5820
CCTTTAAGAC CTTTCTTTTT TTTACGAGAA AAAAGAAACA AAAAAACCTG CCCTCTGCCA 5880
CCTCAGCAAA GGGGGGTTTT GCTCTCGTGC TCGTTAAAAA ATCAGCAAGG GACAGGTAGT 5940
ATTTTTGAG AAGATCACTC AAAAAATCTC CACCTTTAAA CCCTTGCCAA TTTTATTTC 6000
TCGGTTTTG TCTAGCTTAC CGAAAGCCAG ACTCAGCAAG AATAAAATTT TTATTGTCTT 6060
TCGTATCTT TATTCAAGCA TCGCGCCCGA TTGCTGAACA GATTAATAAT AGATTTAGC 6180
TTTTTATTG TTGAAAAAAG CTAATCAAAT TGTTGTCGGG ATCAATTACT GCAAAGTCTC 6240
GTTCATCCCA CCACTGATCT TTTAATGATG TATTGGGGTG CAAAATGCC AAAGGCTTAA 6300
TATGTTGATA TAATTCACTA ATTCCCTCTA CTTCAATGCG GCAACTAGCA GTACCAGCAA 6360
TAAACGACTC CGCACCTGTA CAAACGGTG AATCATTACT ACGAGAGCGC CAGCCTTCAT 6420
CACTTGCCCTC CCATAGATGA ATCCGAACCT CATTACACAT TAGAACTGCG AATCCATCTT 6480
CATGGTGAAC CAAAGTGAAA CCTAGTTAT CGCAATAAAA ACCTATACTC TTTTAATAT 6540
CCCCGACTGG CAATGCCGGG ATAGACTGTA ACATTCTCAC GCATAAAATC CCCTTCATT 6600
TTCTAATGTA AATCTATTAC CTTATTATTA ATTCAATTG CTCATAATTA ATCCTTTTC 6660
TTATTACGCA AAATGGCCCG ATTAAAGCAC ACCCTTTATT CCGTTAATGC GCCATGACAG 6720
CCATGATAAT TACTAATACT AGGAGAAGTT AATAAAATACG TAACCAACAT GATTAACAAT 6780

120

TATTAGAGGT CATCGTTCAA AATGGTATGC GTTTGACAC ATCCACTATA TATCCGTGTC 6840
GTTCTGTCCA CTCCTGAATC CCATTCCAGA AATTCTCTAG CGATTCCAGA AGTTTCTCAG 6900
AGTCGGAAAG TTGACCAGAC ATTACGAACG GGCACAGATG GTCATAACCT GAAGGAAGAT 6960
CTGATTGCTT AACTGCTTCA GTTAAGACCG AAGCGCTCGT CGTATAACAG ATGCGATGAT 7020
GCAGACCAAT CAACATGGCA CCTGCCATTG CTACCTGTAC AGTCAAGGAT GGTAGAAATG 7080
TTGTCGGTCC TTGCACACGA ATATTACGCC ATTTGCCTGC ATATTCAAAC AGCTCTTCTA 7140
CGATAAGGGC ACAAAATCGCA TCGTGGAACG TTTGGGCTTC TACCGATTG GCAGTTGAT 7200
ACACTTTCTC TAAGTATCCA CCTGAATCAT AAATCGGCAA AATAGAGAAA AATTGACCAT 7260
GTGTAAGCGG CCAATCTGAT TCCACCTGAG ATGCATAATC TAGTAGAATC TCTTCGCTAT 7320
CAAAATTACAC TTCCACCTTC CACTCACCGG TTGTCCATTG ATGGCTGAAC TCTGCTTCCT 7380
CTGTTGACAT GACACACATC ATCTCAATAT CCGAATAGGG CCCATCAGTC TGACGACCAA 7440
GAGAGCCATA AACACCAATA GCCTTAACAT CATCCCCATA TTTATCCAAT ATTGTTCCCT 7500
TAATTTCATG AACAATCTTC ATTCTTCTT CTCTAGTCAT TATTATTGGT CCATTCACTA 7560
TTCTCATTCC CTTTCAGAT AATTTAGAT TTGCTTTCT AAATAAGAAT ATTTGGAGAG 7620
CACCGTTCTT ATTCAAGCTAT TAATAACTCG TCTTCCTAAG CATCCTCAA TCCTTTAAT 7680
AACAAATTATA GCATCTAATC TTCAACAAAC TGGCCCGTTT GTTGAACATC TCTTTAATAA 7740
AAATAATTGTTT CCGTTCCCAA TTCCACATTG CAATAATAGA AAATCCATCT TCATCGGCTT 7800
TTTCGTCATC ATCTGTATGA ATCAAATCGC CTTCTCTGT GTCATCAAGG TTTAATTGTTT 7860
TATGTATTTC TTTTAACAAA CCACCATAGG AGATTAACCT TTTACGGTGT AAACCTTCCT 7920
CCAAATCAGA CAAACGTTTC AAATTCTTT CTTCATCATC GGTCAAAAA TCCGTATCCT 7980
TTACAGGATA TTTGCAGTT TCGTCAATTG CCGATTGTAT ATCCGATTG TATTATTGTTT 8040
TCGGTCGAAT CATTGAACT TTTACATTTG GATCATAGTC TAATTCATT GCCTTTTCC 8100
AAAATTGAAT CCATTGTTT TGATTCACGT AGTTTCTGT ATTCTTAAAA TAAGTTGGTT 8160
CCACACATAC CAATACATGC ATGTGCTGAT TATAAGAATT ATCTTTATTA TTTATTGTCA 8220
CTTCCGTTGC ACGCATAAAA CCAACAAGAT TTTTATTAAT TTTTTATAT TGCATCATTC 8280
GGCGAAATCC TTGAGCCATA TCTGACAAAC TCTTATTTAA TTCTTCGCCA TCATAAACAT 8340
TTTTAACTGT TAATGTGAGA AACAAACCAAC GAACTGTTGG CTTTTGTTA ATAACCTCAG 8400
CAACAAACCTT TTGTGACTGA ATGCCATGTT TCATTGCTCT CCTCCAGTTG CACATTGGAC 8460
AAAGCCTGGA TTTACAAAAC CACACTCGAT ACAACTTTCT TTGCGCTGTT TCACGATTGTT 8520

121

GTTTATACTC TAATATTTCA GCACAATCTT TTACTCTTC AGCCTTTA AATTCAAGAA 8580
TATGCAGAAG TTCAAAGTAA TCAACATTAG CGATTTCTT TTCTCTCCAT GGTCTCACTT 8640
TTCCACTTT TGTCTTGTCC ACTAAAACCC TTGATTTTC ATCTGAATAA ATGCTACTAT 8700
TAGGACACAT AATATTAAGA GAAACCCCCA TCTATTTAGT TATTTGTTA GTCACTTATA 8760
ACTTTAACAG ATGGGGTTTT TCTGTGCAAC CAATTTAAG GGTTTCAAT ACTTTAAAAC 8820
ACATACATAC CAACACTTCA ACGCACCTT CAGCAACTAA AATAAAAATG ACGTTATTTC 8880
TATATGTATC AAGATAAGAA AGAACAAAGTT CAAAACCATC AAAAAAAGAC ACCTTTTCAG 8940
GTGCTTTTT TATTTTATAA ACTCATCCC TGATCTCGAC TTCGTTCTT TTTTACCTCT 9000
CGGTTATGAG TTAGTTCAAA TTCGTTCTT TTAGGTTCTA AATCGTGTAA TTCTTGGAAAT 9060
TGTGCTGTTT TATCCTTTAC CTTGTCTACA AACCCCTTAA AAACGTTTTT AAAGGCTTTT 9120
AAGCCGTCTG TACGTTCCCTT AAGG 9144

(2) INFORMATION FOR SEQ ID NO:80:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 303 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:80:

GGGCCGGTCG AGGTGGACAA GGTGCAGGTG CAGCCGCTGC TGCTGCGGGC GGCGCAGGTC 60
AAGGTGGTA TGGGGGTTTA GGTTCACAAAG GGGCCGGACG TGGTGGCCTT GGTGGTCAGG 120
GTGCTGGCGC GGCAGCCGCT GCGGCAGCTG GTGGTGCTGG TCAGGGCGGT CTTGGCTCAC 180
AAGGGGCCGG TCAAGGCGCT GGTGCAGCAG CAGCTGCCGC TGGCGGTGCA GGCCAAGGTG 240
GATATGGTGG CTTAGGGTCA CAAGGGCCGG GGCAAGGTGG TTACGGCGGT CTCGGATCAC 300
AAG 303

(2) INFORMATION FOR SEQ ID NO:81:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 303 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

122

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:81:

GGGCCGGGCA AGGTGGTTAC GGCGGTCTCG GATCACAAAGG GGCGGACGT GGTGGCCTTG 60
GTGGTCAGGG TGCTGGCGCG GCAGCCGCTG CGGCAGCTGG TGGTGCTGGT CAGGGCGGTC 120
TTGGCTCACA AGGGGCCGGT CAAGGCCTG GTGCAGCAGC AGCTGCCGCT GGCGGTGCAG 180
GCCAAGGTGG ATATGGTGGC TTAGGGTCAC AAGGGGCCGG TCGAGGTGGA CAAGGTGCAG 240
GTGCAGCCGC TGCTGCTGCG GGCGCGCAG GTCAAGGTGG GTATGGGGT TTAGGTTCAC 300
AAG 303

(2) INFORMATION FOR SEQ ID NO:82:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 303 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:82:

TCTCAGGGTG CTGGCCAGGG TGGCTATGGT GGCTGGGAT CTCAAGGCAGC TGGTCGCGGT 60
GGCCTGGGTG GCCAGGGTGC AGGTGCTGCT GCTGCTGCAG CTGCTGGTGG TGCAGGTCAG 120
GGTGGTCTGG GATCTCAGGG CGCAGGTCAA GGTGCTGGTGG CAGCTGCCGC GGCAAGCTGGT 180
GGCGCGGGTC AAGGTGGCTA CGGCGGTTA GGATCTCAAG GTGCGGGTCG CGGTGGTCAG 240
GGCGCTGGTG CAGCAGCGGC AGCAGCAGGT GGCGCTGGCC AAGGTGGTTA CGGTGGTCTT 300
GGA 303

(2) INFORMATION FOR SEQ ID NO:83:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 357 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:83:

GGGCCATCCG GCCCAGGTTC TGCGGCAGCG GCAGCAGCGG GCCCAGGGCA GCAGGGGCCG 60
GGCGGTTACG GTCCGGGTCA GCAAGGCCCA GGTGGCTACG GCCCAGGCCA ACAGGGGCCA 120
TCTGGTCCGG GTAGCGCTGC GGCTGCTGCT GCTGCGGCAG GTCCAGGCCG CTACGGGCCG 180

GGCCAACAAG	GTCCGGCGG	CTATGGTCCA	GGTCAACAGG	GGCCGAGCGG	TCCAGGTTCC	240
GCAGCAGCAG	CGGCTGCGGC	GGCAGCGGGT	CCAGGTGGTT	ACGGGCCAGG	CCAGCAGGGT	300
CCGGGTGGCT	ATGGCCCAGG	CCAGCAAGGT	CCGGGTGGTT	ACGGTCCAGG	TCAGCAG	357

(2) INFORMATION FOR SEQ ID NO:84:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:84:

GATCTCAAGG AGCCGGTCAA GGTGGTTACG GAGGTCTGG 39

(2) INFORMATION FOR SEQ ID NO:85:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:85:

GATCCCAGAC CTCCGTAACC ACCTTGACCG GCTCCTTGA 39

(2) INFORMATION FOR SEQ ID NO:86:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 13 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:86:

Ser Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
1 5 10

(2) INFORMATION FOR SEQ ID NO:87:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 93 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:87:

GATCTCAAGG TGCTGGACGT GGTGGTCTTG GTGGTCAGGG TGCCGGTGCC GCCGCTGCCG 60
CCGCCGCTGG TGGTGCTGGA CAAGGTGGTT TGG 93

(2) INFORMATION FOR SEQ ID NO:88:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 93 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:88:

GATCCCAAAC CACCTTGTCC AGCACCACCA GCGGCGGCCGG CAGCGGCCGG ACCGGCACCC 60
TGACCACCAA GACCACCACG TCCAGCACCT TGA 93

(2) INFORMATION FOR SEQ ID NO:89:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 31 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:89:

Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala
1 5 10 15
Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly
20 25 30

(2) INFORMATION FOR SEQ ID NO:90:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 81 base pairs

125

- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:90:

GATCTCAGGG AGCTGGTCAA GGTGCCGGTG CTGCTGCCGC TGCTGCCGGA GGTGCCGGTC 60
AGGGTGGATA CGGTGGACTT G 81

(2) INFORMATION FOR SEQ ID NO:91:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 81 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:91:

GATCCAAGTC CACCGTATCC ACCCTGACCG GCACCTCCGG CAGCAGCGGC AGCAGCACCG 60
GCACCTTGAC CAGCTCCCTG A 81

(2) INFORMATION FOR SEQ ID NO:92:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 27 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:92:

Ser Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Gly
1 5 10 15

Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
20 25

(2) INFORMATION FOR SEQ ID NO:93:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 90 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

126

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:93:

GATCTCAGGG TGCTGGTAGA GGTGGACAAG GTGCCGGAGC TGCCGCTGCC GCTGCCGGTG 60
GTGCTGGTCA AGGAGGTTAC GGTGGTCTTG 90

(2) INFORMATION FOR SEQ ID NO:94:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 90 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:94:

GATCCAAGAC CACCGTAACC TCCTTGACCA GCACCACCGG CAGCGGCAGC GGCAGCTCCG 60
GCACCTTGTC CACCTCTACC AGCACCCCTGA 90

(2) INFORMATION FOR SEQ ID NO:95:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:95:

Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala
1 5 10 15
Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly
20 25 30

(2) INFORMATION FOR SEQ ID NO:96:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 588 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:96:

ATGCATTGTC TCCACATTGT ATGCTTCAA GATTCTGGTG GGAATACTGC TGATAGCCTA 60
 ACGTTCATGA TCAAAATTAA ACTGTTCTAA CCCCTACTTG ACAGCAATAT ATAAACAGAA 120
 GGAAGCTGCC CTGTCTTAAA CCTTTTTTT TATCATCATT ATTAGCTTAC TTTCATAATT 180
 GCGACTGGTT CCAATTGACA AGCTTTGAT TTTAACGACT TTTAACGACA ACTTGAGAAG 240
 ATCAAAAAAC AACTAATTAT TCGAAACGAT GAGATTCCT TCAATTTTA CTGCAGTTT 300
 ATTCCGAGCA TCCTCCGCAT TAGCTGCTCC AGTCAACACT ACAACAGAAG ATGAAACGGC 360
 ACAAAATTCCG GCTGAAGCTG TCATCGGTTA CTCAGATTAA GAAGGGGATT TCGATGTTGC 420
 TGTTTGCCA TTTCCAACA GCACAAATAA CGGGTTATTG TTTATAAATA CTACTATTGC 480
 CAGCATTGCT GCTAAAGAAG AAGGGGTATC TCTCGAGAAA AGAGAGGCTG AAGCTTACGT 540
 AGAATTCCCT AGGGCGGCCG CGAATTAATT CGCCTTAGAC ATGACTGT 588

(2) INFORMATION FOR SEQ ID NO:97:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 93 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: unknown
- (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:97:

Met	Arg	Phe	Pro	Ser	Ile	Phe	Thr	Ala	Val	Leu	Phe	Ala	Ala	Ser	Ser
1					5					10				15	
Ala	Leu	Ala	Ala	Pro	Val	Asn	Thr	Thr	Thr	Glu	Asp	Glu	Thr	Ala	Gln
					20				25				30		
Ile	Pro	Ala	Glu	Ala	Val	Ile	Gly	Tyr	Ser	Asp	Leu	Glu	Gly	Asp	Phe
					35			40			45				
Asp	Val	Ala	Val	Leu	Pro	Phe	Ser	Asn	Ser	Thr	Asn	Asn	Gly	Leu	Leu
				50				55			60				
Phe	Ile	Asn	Thr	Thr	Ile	Ala	Ser	Ile	Ala	Ala	Lys	Glu	Glu	Gly	Val
					65			70		75			80		
Ser	Leu	Glu	Lys	Arg	Glu	Ala	Glu	Ala	Tyr	Val	Glu	Phe			
					85				90						

(2) INFORMATION FOR SEQ ID NO:98:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 base pairs

128

- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:98:

CAACTAATTA TTGAAACGA TGAGATTCC

30

(2) INFORMATION FOR SEQ ID NO:99:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 23 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:99:

CTGAGGAAACA GTCATGTCTA AGG

23

(2) INFORMATION FOR SEQ ID NO:100:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 30 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:100:

GGAAATCTCA TCGTTTCGAA TAATTAGTTG

30

(2) INFORMATION FOR SEQ ID NO:101:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 23 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:101:

GAAACGCAAA TGGGGAAACA ACC

23

(2) INFORMATION FOR SEQ ID NO:102:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 9 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS: unknown
 - (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:102:

Met Gly Ser His His His His His His
1 5

(2) INFORMATION FOR SEQ ID NO:103:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 32 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:103:

AATTATGGGA TCCCATCACC ATCACCATCA CT

32

(2) INFORMATION FOR SEQ ID NO:104:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 32 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:104:

AATTAGTGAT GGTGATGGTG ATGGGATCCC AT

32

(2) INFORMATION FOR SEQ ID NO:105:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 6 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS: unknown
 - (D) TOPOLOGY: unknown

(ii) MOLECULE TYPE: peptide

130

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:105:

Phe Gly Ser Gln Gly Ala
1 5

(2) INFORMATION FOR SEQ ID NO:106:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 23 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:106:

AATTCTGGATC CCAGGGTGCT TAA

23

(2) INFORMATION FOR SEQ ID NO:107:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 23 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:107:

GGCCTTAAGC ACCCTGGGAT CCG

23

We claim:

1. A novel synthetic spider dragline variant protein produced by a process comprising the steps of:

- (i) designing a DNA monomer sequence of between about 50 bp and 1000 bp which codes for an polypeptide monomer consisting of a variant of a consensus sequence derived from the fiber forming regions of spider dragline protein;
- 10 (ii) assembling said DNA monomer;
- (iii) polymerizing said DNA monomer to form a synthetic gene encoding a full length silk variant protein wherein said synthetic gene does not encode any portion of the *Nephila clavipes* genome;
- 15 (iv) transforming a suitable host cell with a vector containing said synthetic gene;
- (v) expressing said synthetic gene whereby the protein encoded by said gene is produced at levels between 1 mg and 300 mg of full-length protein per gram of cell mass; and
- 20 (vi) recovering said protein in a useful form.

2. A composition consisting essentially of the nucleic acid sequence:

GGGCCGGTCG	AGGTGGACAA	GGTGCAGGTG	CAGCCGCTGC	TGCTGCGGGC	GGCGCAGGTC	60	
AAGGTGGGTA	TGGGGGTTTA	GGTCACAAG	GGGCCGGACG	TGGTGGCCTT	GGTGGTCAGG	120	
30	GTGCTGGCGC	GGCAGCCGCT	GCAGCAGCTG	TGGTGCTGG	TCAGGGCGGT	180	
	AAGGGGCCGG	TCAAGGCGCT	GGTGCAGCAG	CAGCTGCCGC	TGGCGGTGCA	GGCCAAGGTG	240
	GATATGGTGG	CTTAGGGTCA	CAAGGGCCG	GGCAAGGTGG	TTACGGCGGT	CTCGGATCAC	300
	AAG						303

wherein said sequence designated SEQ ID NO.:80 encodes 35 the DP-1A.9 amino acid monomer.

3. A composition consisting essentially of a nucleic acid sequence which when polymerized encodes a spider silk variant protein comprising from 1 to 16 tandem repeats of the DP-1A.9 amino acid monomer.

5 4. A composition consisting of from 1 to 16 tandem repeats of the nucleic acid sequence of Claim 2.

5. A composition consisting essentially of the nucleic acid sequence:

GGGCCGGGCA AGGTGGTTAC GGCGGTCTCG GATCACAAAGG GGCCGGACGT GGTGGCCTTG 60
10 GTGGTCAGGG TGCTGGCGCG GCAGCCGCTG CGGCAGCTGG TGGTGCTGGT CAGGGCGGTC 120
TTGGCTCACA AGGGGCCGGT CAAGGCCTG GTGCAGCAGC AGCTGCCGCT GGCGGTGCAG 180
GCCAAGGTGG ATATGGTGGC TTAGGGTCAC AAGGGGCCGG TCGAGGTGGA CAAGGTGCAG 240
GTGCAGCCGC TGCTGCTGCG GGCGCGCAG GTCAAGGTGG GTATGGGGT TTAGGTTCAC 300
AAG 303

15 wherein said sequence designated SEQ ID NO.:81 encodes the DP-1B.9 amino acid monomer.

6. A composition consisting essentially of a nucleic acid sequence which when polymerized encodes a spider silk variant protein comprising from 1 to 16 tandem repeats of the DP-1B.9 amino acid monomer.

7. A composition consisting of from 1 to 16 tandem repeats of the nucleic acid sequence of Claim 5.

8. A composition consisting essentially of the nucleic acid sequence:

25 TCTCAGGGTG CTGGCCAGGG TGGCTATGGT GGCCTGGGAT CTCAAGGCAGC TGGTCGCGGT 60
GGCCTGGGTG GCCCAGGGTGC AGGTGCTGCT GCTGCTGCGG CTGCTGGTGG TGCAGGTCAG 120
GGTGGTCTGG GATCTCAGGG CGCAGGTCAA GGTGCTGGTGG CAGCTGCCGC GGCAAGCTGGT 180
GGCGCGGGTC AAGGTGGCTA CGGCGGGTTA GGATCTCAAG GTGCGGGTGC CGGTGGTCAG 240
GGCGCTGGTG CAGCAGCGGC AGCAGCAGGT GGCGCTGGCC AAGGTGGTTA CGGTGGTCTT 300
30 GGA 303

wherein said sequence designated SEQ ID NO.:82 encodes the DP-1B.16 amino acid monomer.

9. A composition consisting essentially of a nucleic acid sequence which when polymerized encodes a

spider silk variant protein comprising from 1 to 16 tandem repeats of the DP-1B.16 amino acid monomer.

10. A composition consisting of from 1 to 16 tandem repeats of the nucleic acid sequence of Claim 8.

5 11. A composition consisting essentially of the nucleic acid sequence:

GGGCATCCG GCCCAGGTTTC TGCGGCAGCG GCAGCAGCGG GCCCAGGGCA GCAGGGGCCG 60

GGCGGTTACG GTCCGGGTCA GCAAGGCCA GGTGGCTACG GCCCAGGCCA ACAGGGGCCA 120

TCTGGTCCGG GTAGCGCTGC GGCTGCTGCT GCTGCGGCAG GTCCAGGCCG CTACGGGCCG 180

10 GGCCAACAAG GTCCGGGCCGG CTATGGTCCA GGTCAACAGG GGCGAGCGG TCCAGGTTCC 240

GCAGCAGCAG CGGCTGCGGC GGCAAGCGGT CCAGGTGGTT ACGGGCCAGG CCAGCAGGGT 300

CCGGGTGGCT ATGGCCCAGG CCAGCAAGGT CCGGGTGGTT ACGGTCCAGG TCAGCAG 357

wherein said sequence designated SEQ ID NO.:83 encodes the DP-2A amino acid monomer.

15 12. A composition consisting essentially of a nucleic acid sequence which when polymerized encodes a spider silk variant protein comprising from 1 to 16 tandem repeats of the DP-2A amino acid monomer.

20 13. A composition consisting of from 1 to 16 tandem repeats of the nucleic acid sequence of Claim 11.

25 14. A plasmid comprising the compositions of Claims 3, 6, 9, or 12 operably and expressibly linked to a suitable promoter wherein said plasmid is capable of transforming a host cell for the expression of a spider silk variant protein at levels between 1 mg and 300 mg of full-length protein per gram of cell mass.

30 15. A plasmid as recited in Claim 14 wherein said compositions are flanked on either the 5' end or the 3' end by a DNA fragment encoding a series of between 4 and 20 histidine residues.

16. A transformed host cell comprising the plasmid of Claims 14 or 15 capable of expressing a spider silk variant protein at levels between 1 mg and 300 mg of full-length protein per gram of cell mass.

17. A host cell as recited in Claim 16 wherein said host cell is selected from the group consisting of *E. coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Pichia pastoris*, *Aspergillus* 5 *sp*, and *Streptomyces* *sp*.

18. A host cell transformed with a plasmid comprising compositions of Claims 3, 8, 9, or 12 said host cell is capable of secreting spider silk variant protein into the cell growth media.

10 19. The transformed *E. coli* host FP3350 identified by the ATCC number ATCC 69328.

20. The transformed *Bacillus subtilis* host FP2193, identified by the ATCC number ATCC 69327.

15 21. A universal expression vector pFP204, useful for the expression of spider silk variant proteins, said vector being devoid of any synthetic spider silk variant DNA, wherein said expression vector is contained in a bacterial strain identified by the ATCC number ATCC 69326.

20 22. A method for the production of a synthetic spider dragline variant protein comprising the steps of:

- (i) designing a DNA monomer sequence of between about 50 bp and 1000 bp which codes for an polypeptide monomer 25 consisting of a variant of a consensus sequence derived from the fiber forming regions of spider dragline protein;
- (ii) assembling said DNA monomer;
- (iii) polymerizing said DNA monomer to form a synthetic gene encoding a full length silk variant protein;
- (iv) transforming a suitable host cell with a vector containing said synthetic gene;

5 (v) expressing said synthetic gene whereby the protein encoded by said gene is produced at levels between 1 mg and 300 mg of full-length protein per gram of cell mass; and

(vi) recovering said protein in a useful form.

23. A method for the production of a synthetic spider dragline variant protein comprising the steps of:

10 (i) designing a DNA monomer sequence of between about 50 bp and 1000 bp which codes for an polypeptide monomer consisting of a variant of a consensus sequence derived from the fiber forming regions of spider dragline protein;

15 (ii) assembling said DNA monomer;

(iii) polymerizing said DNA monomer to form a synthetic gene encoding a full length silk variant protein;

20 (iv) transforming a suitable host cell with a vector containing said synthetic gene;

(v) expressing said synthetic gene whereby the protein encoded by said gene is secreted into the extracellular medium;

25 (vi) recovering said protein in a useful form.

24. A spider dragline variant protein as recited in Claim 1 wherein said full length variant protein is defined by the formula:

[ACQGGYGGLGXQGAGRGGLGGQGAGAnGG]z

wherein X=S, G or N; n=0-7 and z=1-75, and wherein:

30 (a) when n=0 the sequence encompassing

35 AGRGGLGGQGAGAnGG is deleted;

(b) deletions other than poly-alanine sequence will encompass integral multiples of three consecutive residues;

5 (c) the deletion of GYG is accompanied by deletion of GRG in the same repeat; and

(d) a repeat in which the entire poly-alanine sequence is deleted is preceded by a repeat containing six alanine residues; and

10 wherein the full-length protein is not encoded by any portion of the *Nephila clavipes* genome.

25. A spider dragline variant protein as recited in Claim 1 wherein said full length silk variant protein is defined by the formula:

[GPGGYGPGQQGPGGYGPQQGPGGYGPGQQGPGSAn]z

15 wherein n=6-10 and z=1-75 and wherein, excluding the poly-alanine sequence, individual repeats differ from the consensus repeat sequence by deletions of integral multiples of five consecutive residues consisting of one or both of the pentapeptide sequences GPGGY or GPGQ and 20 wherein the full-length protein is not encoded by any portion of the *Nephila clavipes* genome.

1/28

FIG. 1

1		...	QG A GAAAAAA-GG
2	A GQG GYG GLG GQG - --- --- --- - -----		
3	A GQG GYG GLG GQG A --- --- GQG A GAAAAAAAGG		
4	A GQG GYG GLG SQG A GRG --- GQG A GAAAAAA-GG		
5	A GQG GYG GLG SQG A GRG GLG GQG A GAAAAAAAGG		
6	A GQG GYG GLG NQG A GRG --- GQG - --AAAAAAAGG		
7	A GQG GYG GLG SQG A GRG GLG GQG A GAAAAAA-GG		
8	A GQG GYG GLG GQG - --- --- --- - -----		
9	A GQG GYG GLG SQG A GRG GLG GQG A GAAAAAAAGG		
10	A GQG --- GLG GQG A --- --- GQG A GASAAA-GG		
11	A GQG GYG GLG SQG A GRG --- GEG A GAAAAAA-GG		
12	A GQG GYG GLG GQG - --- --- --- - -----		
13	A GQG GYG GLG SQG A GRG GLG GQG A GAAAA---GG		
14	A GQG --- GLG GQG A --- --- GQG A GAAAAAA-GG		
15	A GQG GYG GLG SQG A GRG GLG GQG A GAVAAAAAGG		
16	A GQG GYG GLG SQG A GRG --- GQG A GAAAAAA-GG		
17	A GQR GYG GLG NQG A GRG GLG GQG A GAAAAAAAGG		
18	A GQG GYG GLG NQG A GRG --- GQG - --AAAAA-GG		
19	A GQG GYG GLG SQG A GRG --- GQG A GAAAAAA-VG		
20	A GQE --- GIR GQG - --- --- --- - -----		
21	A GQG GYG GLG SQG S GRG GLG GQG A GAAAAAA-GG		
22	A GQG --- GLG GQG A --- --- GQG A GAAAAAA-GG		
23	V RQG GYG GLG SQG A GRG --- GQG A GAAAAAA-GG		
24	A GQG GYG GLG GQG V GRG GLG GQG A GAAA---GG		
25	A GQG GYG GVG S-- - --- --- G A SAASAAA--		

SEQ. NO. 19

2/28

FIG.2A

"MONOMER":

G	AGRGG---GQGAGAAAAAA-GG SEQ. NO. 20
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQ	-----

FIG.2B

"POLYMER":

-	AGRGG---GQGAGAAAAAA-GG SEQ. NO. 21
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQG	AGRGG---GQGAGAAAAAA-GG
AGQGGYGGLGSQG	AGRGGGLGGQGAGAAAAAAAGG
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG
AGQGGYGGLGSQG	-----
AGQGGYGGLGSQ	-----

3/28

FIG.3A

"MONOMER":

G	-----	SEQ. NO. 22
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQ		

FIG.3B

"POLYMER":

G	-----	SEQ. NO. 23
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	-----	
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	-----	
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	-----	
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	-----	
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	-----	
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	-----	
AGQGGYGGLGSQG	AGRGGLGGQGAGAAAAAAAGG	
AGQGG---LGSQG	A-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	AGR-----GQGAGAAAAAA-GG	
AGQGGYGGLGSQG	-----	

4/28

FIG. 4A

Oligonucleotide L

Oligonucleotide M1

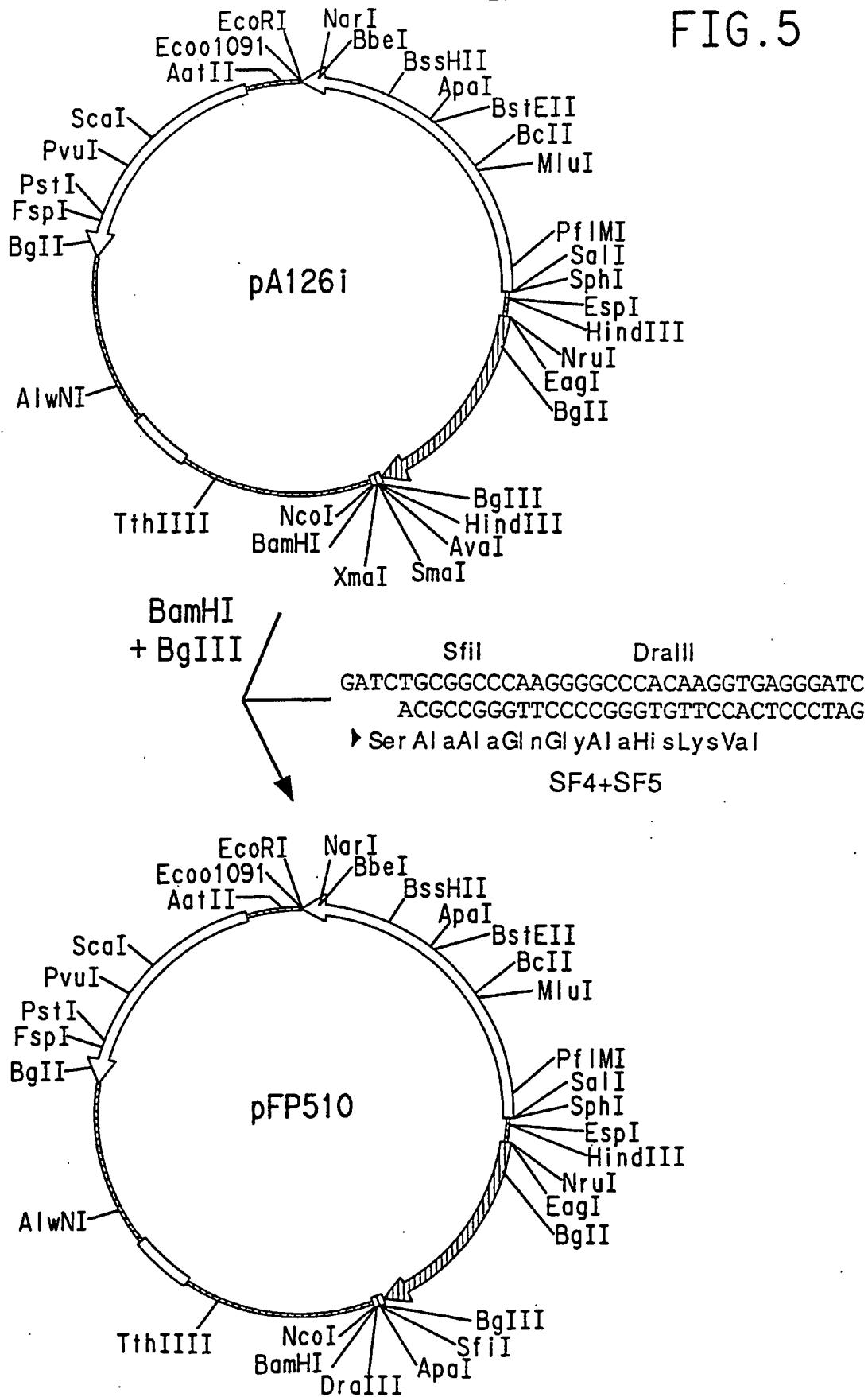
Oljaguncleotide M2

Oligonucleotide S

GGCCCCGGCCAGGCTGGCTTGCCGCTCGCTCGGTCAACAG
 TTCCCCGGCCCGTTCCTCCAAATGCCCGCAGGCTCTAGTG
 ▶ GI yAI aGI yGI nGI yGI yLoughysor GI n

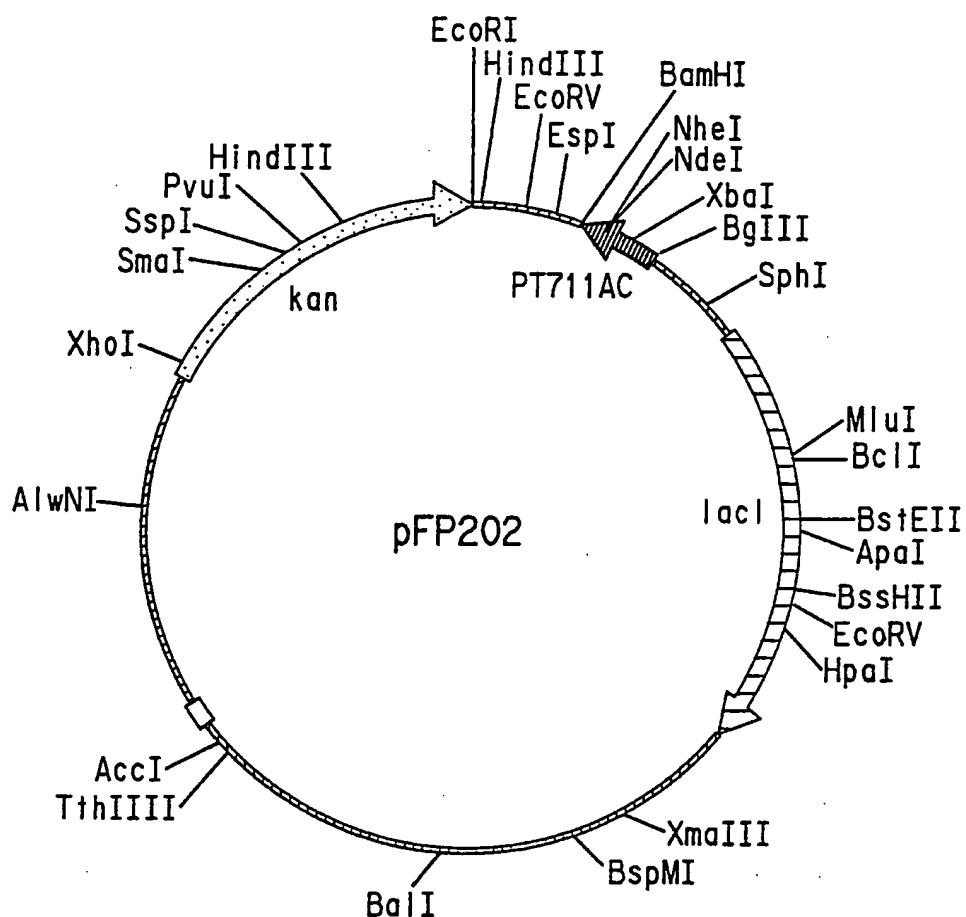
5/28

FIG.5



6/28

FIG. 6



BamET

SEQ. NO. 39

... GGA TCC CAT CAC CAT CAC TCT AGA TCC GGC TGC TAA
... Gly Ser His His His His His Ser Arg Ser Gly Cys END

SEQ. NO. 40

9/28

FIG. 8

SEQ. NO. 59

...PGGY GPGQQ GPGGY GPGQQ GP--SGPGS AAAAAAAA
GPGGY GPGQQ GPGGY GPGQQ GPGRY GPGQQ GP--SGPGS AAAAAA---
----. GSGQQ GPGGY GPGQQ GPGGY GQGQQ GP--SGPGS AAAASAAASA ESGQQ
GPGGY GPGQQ GPGGY GPGQQ GPGGY GPGQQ GP--SGPGS AAAAAAAAS-
---- GPGQQ GPGGY GPGQQ GPGGY GPGQQ GP--SGPGS AAAAAAAAS-
---- GPGQQ GPGGY GPGQQ GPGGY GPGQQ GL--SGPGS AAAAAAA---
---- ---- GPGQQ GPGGY GPGQQ GP--SGPGS AAAAAAAA-
---- GPGGY GPGQQ GPGGY GPGQQ GP--SGAGS AAAAAAA---
---- GPGQQ GLGGY GPGQQ GPGGY GPGQQ GPGGYGPJS ASAAAAAA---
---- ---- GPGQQ GPGGY GPGQQ GP--SGPGS ASAAAAAAA
---- GPGGY GPGQQ GPGGY APGQQ GP--SGPGS ASAAAASAAA
---- GPGGY GPGQQ GPGGY APGQQ GP--SGPGS AAAAAAASA-
---- ---- GPGGY GPAQQ GP--SGPGI AASAASA---
---- ---- GPGGY GPAQQ GPAGYGPJS AVAASA---
---- ---- ---- GA GSAGYGPJS QASAAAS---

10/28

FIG. 9A

"MONOMER": 119 g

SEQ. NO. 60
| GP--SGPGS AAAAAA-----
----. GPGQQ|GPGGY GPGQQ GPGGY GPGQQ|GP--SGPGS AAAAAAAA--
----. GPGGY|GPGQQ GPGGY GPGQQ|GP--SGPGS AAAAAAAA--
GPGGY|GPGQQ GPGGY GPGQQ GPGGY GPGQQ|

FIG. 9B

"POLYMER" •

11/28

FIG.10A

"MONOMER":

S Q G	-----	SEQ. NO. 62
A G Q G G Y G G L G S Q G	A G R G G L G G Q G A G A A A A A A A A A G G	
A G Q G G G -- L G S Q G	A ----- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	A G R G G -- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G		

FIG.10B

"POLYMER":

S Q G	-----	SEQ. NO. 63
A G Q G G Y G G L G S Q G	A G R G G L G G Q G A G A A A A A A A A A G G	
A G Q G G G -- L G S Q G	A ----- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	A G R G G -- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	-----	
A G Q G G Y G G L G S Q G	A G R G G L G G Q G A G A A A A A A A A A G G	
A G Q G G G -- L G S Q G	A ----- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	A G R G G -- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	-----	
A G Q G G Y G G L G S Q G	A G R G G L G G Q G A G A A A A A A A A A G G	
A G Q G G G -- L G S Q G	A ----- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	A G R G G -- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	-----	
A G Q G G Y G G L G S Q G	A G R G G L G G Q G A G A A A A A A A A A G G	
A G Q G G G -- L G S Q G	A ----- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	A G R G G -- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	-----	
A G Q G G Y G G L G S Q G	A G R G G L G G Q G A G A A A A A A A A A G G	
A G Q G G G -- L G S Q G	A ----- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	A G R G G -- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	-----	
A G Q G G Y G G L G S Q G	A G R G G L G G Q G A G A A A A A A A A A G G	
A G Q G G G -- L G S Q G	A ----- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G S Q G	A G R G G -- G Q G A G A A A A A A A A - G G	
A G Q G G Y G G L G		

FIG. 11A

Oligonucleotide 1

5' Ser1Gly1Ala1Gly2Gly3Gly4Gly5Gly6Gly7Gly8Gly9Gly10Gly11Gly12Gly13Gly14Gly15Gly16Gly17Gly18Gly19Gly20Gly10' SEQ. NO. 64
 GATCTCAGGGCTGGCCHGGTGGCTATGGTGGCTGG
 AGTCCCAGGACGGTCCACCGATACCAACGGACCCCTAG
 GATCTCAGGGCTGGCCHGGTGGCTATGGTGGCTGG
 AGTCCCAGGACGGTCCACCGATACCAACGGACCCCTAG
 SEQ. NO. 65
 SEQ. NO. 66

FIG. 11B

Oligonucleotide 2

FIG. 11C

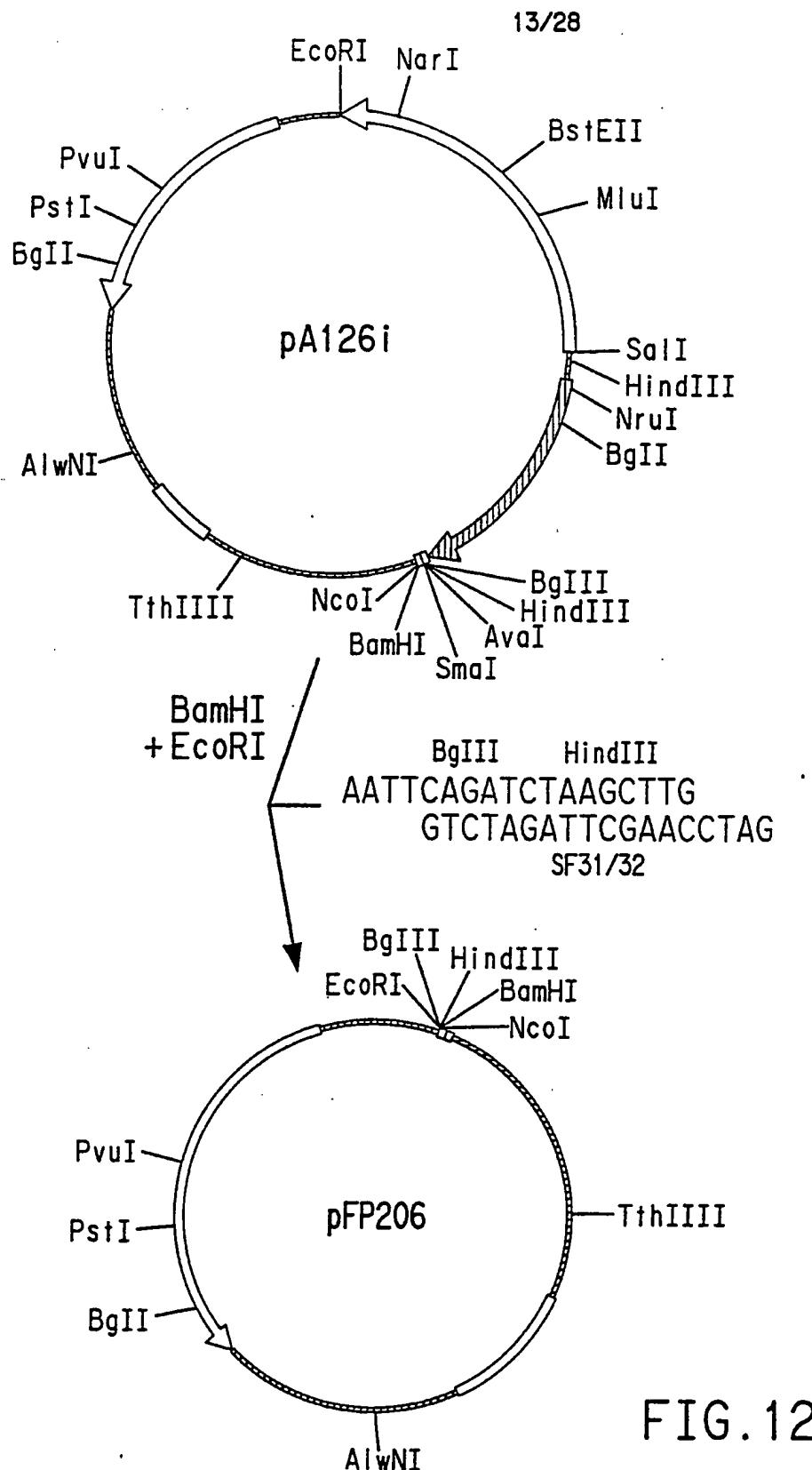
Oligonucleotide 3

FIG. 11D

Oligonucleotide 4

12/28

12/28



14/28

FIG. 13A

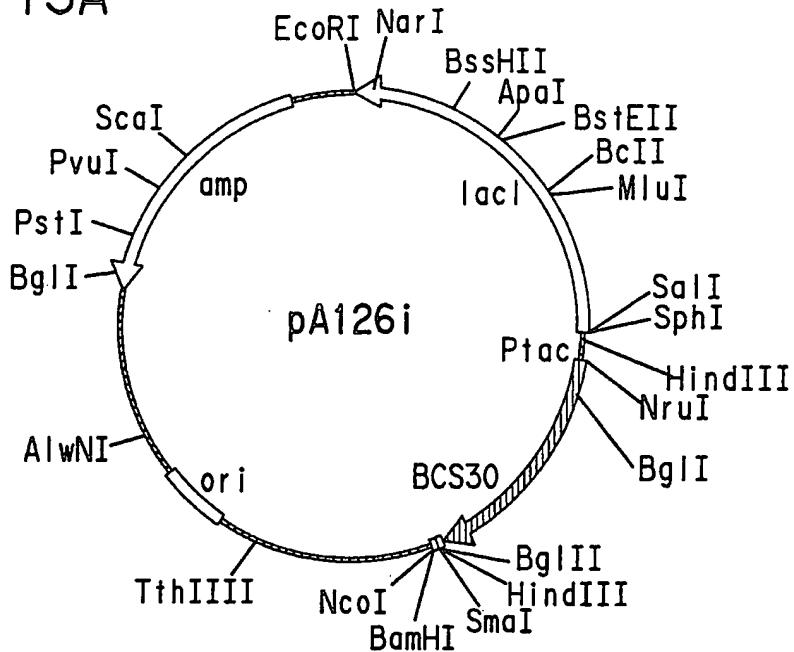


FIG. 13B

EcoRI

SEQ. NO. 78

4909 GAATTCCGGGGATTATGCGTTAACATAAAGTGTAAAGCCTGGGTGCCTA
 4961 ATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTCCAG
 5013 TCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGA

Bbel

Narl

5065 GAGGCGGTTTGCATTGGCGCCAGGGTGGTTTCTTTCAACCAGTGAGA

5117 CGGGCAACAGCTGATTGCCCTCACCGCCTGGCCCTGAGAGAGTTGCAGCAA

5169 GCGGTCCACGCTGGTTGCCAGCAGGCGAAAATCCTGTTGATGGTGGTT

5221 GACGGCGGGATATAACATGAGCTGTCGGTATCGTCGTATCCCACCTACCG

BssHII

5273 AGATATCCGCACCAACGCGCAGCCGGACTCGTAATGGCGCCATTGCGCC

5325 CAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGAACGATGCCCTCA

5377 TTCAGCATTTGCATGGTTGTTGAAAACCGGACATGGCACTCCAGTCGCCCT

5429 CCCGTTCCGCTATCGGCTGAATTGATTGCGAGTGAGATATTATGCCAGCC

ApaI

5481 AGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCG

BstEII

5533 ATTTGCTGGTACCCAATGCGACCAGATGCTCCACGCCAGTCGCGTACCGT

5585 CTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAG

5637 AAATAACGCCGGAACATTAGTGCAGGGCAGCTCCACAGCAATGGCATCCTGG

BclI

MluI

5689 TCATCCAGCGGATAGTTAATGATCAGCCCAGTGACGCGTTGCGCGAGAAGAT

5741 TGTGCACCGCCGCTTACAGGCTTCGACGCCGTTCTACCATCGACAC

5793 CACCACGCTGGCACCCAGTTGATGGCGCGAGATTTAATGCCCGGACAATT

5845 TGCAGGGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAACAGAACG

SUBSTITUTE SHEET (RULE 26)

15/28

FIG.13C

5897 ACTGTTGCCCGCCAGTTGTGCCACGCCGGTGGGAATGTAATTCAAGCTC
 5949 CGCCATGCCGCTTCCACTTTCCCGCTTTCGAGAAACGTGGCTGGCC
 6001 TGGTTACCAACGCCGGAAACGGTCTGATAAGAGACACCAGCATACTCTGCGA
 6053 CATCGTATAACGTTACTGGTTACATTCACCACCCCTGAATTGACTCTCTC
 6105 CGGGCGCTATCATGCCATACCGCGAAAGGTTTGCGCCATTGATGGTGTCA
 6157 ACCTTGCAGAGCTGCGCTTATTATTATCCGCCGGGAGAAAATATTCCGTG
 Sall SphI
 6209 GATCTAACGGGATGCGTTATGTTGAAGTGAGACCGGTCGACGCATGCCAGGA
 HindIII
 6261 CAACTCTGGTCCCGTAACGTGCTGAGCCCGCCAAGCTTACTCCCCATCCC
 6313 CCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGAT
 6365 ACAAAATTCACACAGGAAACAGGATCACTAAGGAGGTTAAATATGGCTACT
 NruI
 6417 GTTATAGATCCGTCTGTCGCGACGGCCGTTCGTCGAATGGCTCGGTTGCCA
 6469 ATATCAATGCGATCAAGTCGGCGCTCTGGAGTCCGGTTACGCAGTCAGA
 BglII
 6521 CGTTGCCTATTGGCCCTATAACGGCACCGGCCCTTATGATGGCAAGGGCAAG
 6573 GTGGAAGATTGCGCCTCTGGCGACGCTTACCCGAAACGATCCATATCG
 6625 TTGCGCGTAAGGATGCAAACATCAAATCGGTCGCAGACCTGAAAGGCAAGCG
 6677 CGTTTCGCTGGATGAGCCGGTTCTGGCACCATCGTCGATGCCGTATCGTT
 6729 CTTGAAGCCTACGGCCTACCGGAAGACGATATCAAGGCTGAACACCTGAAGC
 6781 CGGGACCGGCAGGCCAGAGGGCTGAAAGATGGTGCCTGGACGCCATTCTT
 6833 TGTGGGCGGCTATCCGACGGCGCAATCTCGGAACCTGCCATCTGAACGGT
 6885 ATTTCGCTCGTTCCGATCTCGGGCCCGAAGCCGACAAGATTCTGGAGAAAT
 6937 ATTCCCTCTCTCGAAGGATGTGGTCTGCCGGAGCCTATAAGGACGTGGC
 6989 GGAAACACCGACCCCTGCCGTTGCCGCACAGTGGGTGACGAGCGCCAAGCAG
 7041 CGGGACGACCTCATATAACATCACCAAGGCTGGTTCTCGAAACCGGGTG
 BglIII HindIII SmaI BamHI Ncol
 7093 CTGGTAGATCTAACGCTCCCGGGGATCCTAGCTAGCTAGCCATGGCATCACA
 7145 GTATCGTGTGACAGAGGGCAGGGAGTGGGACAAAATTGAAATCAAATAATGA
 7197 TTTTATTTGACTGATAGTGCACCTGTTGCAACAAATTGATAAGCAATG
 7249 CTTTTTATAATGCCAACTTAGTATAAAAAAGCTGAACGAGAAACGTAAAT
 7301 GATATAAAATATCAATATATTAAATTAGATTTGCATAAAAAACAGACTACAT
 7353 AATACTGTAAAACACAATATGCAGTCACTATGAATCAACTACTTAGATGG
 7405 TATTAGTGCACCTGTAACAGAGCATTAGCGCAAGGTGATTTGCTTCTTGC
 7457 GCTAATTTTGTGATCAAACCTGTCGCACTCCAGAGAAGCACAAAGCCTCG
 7509 CAATCCAGTGCAAAGCTCTGCCCGCGTTCGGTGATGACGGTGAAAACC
 7561 TCTGACACATGCAGCTCCGGAGACGGTCACAGCTGCTGTAAGCGGATGC
 7613 CGGGAGCAGACAAGCCGTCAGGGCGCGTCAGCGGGTGTGGCGGGTGTGG
 TthIII
 7665 GGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAA
 7717 CTATGCGGCATCAGAGCAGATTGACTGAGAGTGCACCATATGCCGTGTGAA
 7769 ATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCCTCTCCGCTT
 7821 CCTCGCTACTGACTCGCTCGCTCGGTGTTGGCTGCGCGAGCGGTATC
 7873 AGCTCACTCAAAGGCGGTAAACGGTTATCCACAGAATCAGGGGATAACGCA
 7925 GGAAAGAACATGTGAGCAAAGGCCAGCAGGAAACCGTAAAAGG
 7977 CCGCGTTGCTGGCTTTCCATAGGCTCCGCCCGGGCTGACGAGCATACAA
 8029 AAATCGACGCTCAAGTCAGAGGTGGCGAAACCGACAGGACTATAAGATAAC
 8081 CAGGCCTTCCCCCTGGAAGCTCCCTCGTGCCTCTGTTCCGACCCCTGC

16/28

FIG. 13D

8133 CGCTTACCGGATAACCTGTCCGCCTTCTCCCTTCGGGAAGCGTGGCGCTTC
 8185 TCATAGCTCACGCTGTAGGTATCTCAGTCGGTGTAGGTCGTCGCTCCAAG
 8237 CTGGGCTGTGTGCAACCCCCCGTTCAAGCCGACCGCTGCGCCTTATCCG
 8289 GTAACTATCGTCTGAGTCCAACCCGGTAAGACACGACTTATGCCACTGGC
 AlwNI
 8341 AGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCAGGTGCTACA
 8393 GAGTTCTTGAAGTGGTGGCCTAATCAGGCTACACTAGAAGGACAGTATTG
 8445 GTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTC
 8497 TTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTGTTGCAAG
 8549 CAGCAGATTACGCGCAGAAAAAAGGATCTAAGAAGATCCTTGATCTT
 8601 CTACGGGTCTGACGCTCAGTGGAACGAAAACACGTTAAGGGATTGGT
 8653 CATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTAAATTAAAAATGA
 8705 AGTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACC
 8757 AATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTGTCATC
 8809 CATAGTTGCCTGACTCCCCGTGTAGATAACTACGATAACGGGAGGGCTTA
 8861 CCATCTGGCCCCAGTGCTGCAATGATAACCGCGAGACCCACGCTCACCGGCTC
 BglII
 8913 CAGATTTATCAGCAATAAACCAAGCCAGCCAGCGGAAGGGCCGAGCGCAGAAGTGG
 8965 TCCTGCAACTTATCCGCCTCCATCCAGTCATTAAATTGTTGCCGGAAAGCT
 PstI
 9017 AGAGTAAGTAGTCGCCAGTTAATAGTTGCGCAACGTTGGCATTGCTG
 9069 CAGGCATCGTGGTGTACGCTCGTCTGGTATGGCTTCATTCAAGCTCCGG
 9121 TTCCCAACGATCAAGGCGAGTTACATGATCCCCATGTTGTGCAAAAAGCG
 PvuI
 9173 GTTAGCTCTCGGTCTCGATCGTTGTAGAAGTAAGTTGGCCGAGTGT
 9225 TATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATC
 ScaI
 9277 CGTAAGATGCTTTCTGTGACTGGTAGTACTCAACCAAGTCATTCTGAGAA
 9329 TAGTGTATGCGGCACCGAGTTGCTCTTGCCTGGCGTCAACACGGGATAATA
 9381 CCGCGCCACATAGCAGAACTTTAAAGTGCTCATCATTGGAAAACGTTCTTC
 9433 GGGGCGAAAACCTCAAGGATCTTACCGCTGGAGATCCAGTTGATGTAA
 9485 CCCACTCGTGCACCCAACGTGATCTTCAGCATCTTACTTCAACCAGCGTT
 9537 CTGGGTGAGCAAAACAGGAAGGCAAAATGCCGCAAAAAGGGAATAAGGGC
 9589 GACACGGAAATGTGAATACTCATACTCTTCTTTCAATATTATTGAAGC
 9641 ATTTATCAGGGTTATTGTCATGAGCGGATACATATTGAATGTATTAGA
 9693 AAAATAAACAAATAGGGTCCGCGCACATTCCCCGAAAAGTGCCACCTGA
 9745 CGTCTAAGAACCATATTATCATGACATTAACCTATAAAATAGGCGTATC
 9797 ACGAGGCCCTTCGTCTCAA

FIG. 14A

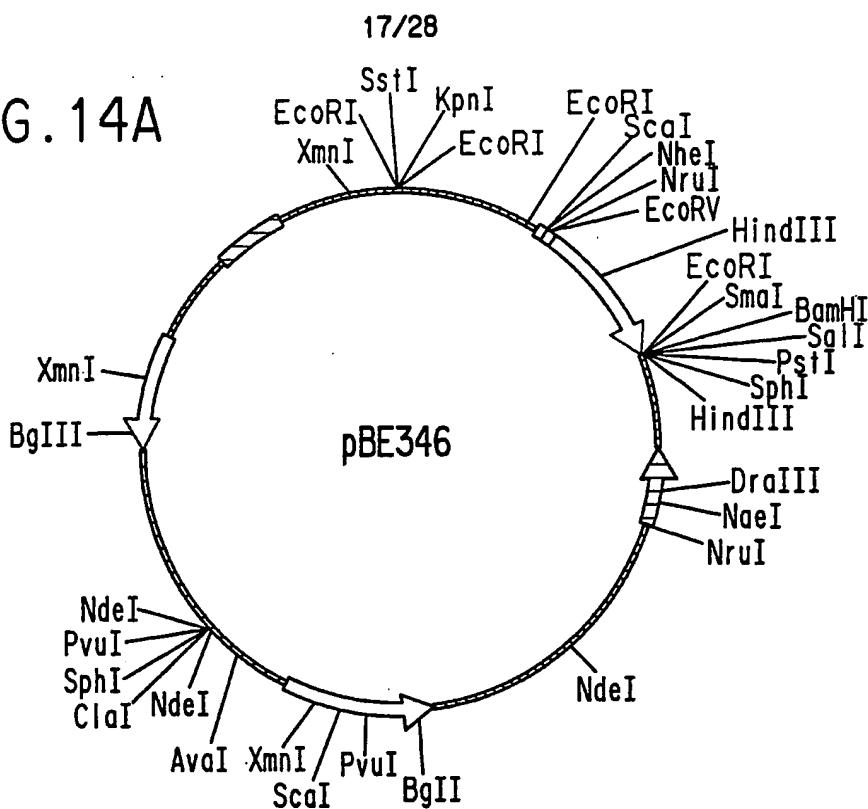


FIG. 14B

SEQ. NO. 79

	SstI	KpnI	EcoRI
1	AATT	CGAGCTCGGTACCCATCGAATT	CCTCAGGAAAAGAACGATGGCTGTC
53	TTATT	AGCGGTTGCAGGCACATT	ATTTGGTCACACACGGGAATGTCGGCA
105	GCCTGTCTATATCCGGTCTGGCTGT	TTTTGGGGCATCAGCTCGGCATTG	
157	GCTGGCGTTTACACCCTCCAGCCGCATCGGCTT	TTGAAGAAATGGGGCTCC	
209	GCCATTATTGTCGGATGGGCATGCTGATGCGGAGCCGTTCTCAGCCTGATT		
261	CAGCCGCCTTGGAAAGTTGAAGGCCAATGGTCGTTGTCGCATATGCCCGA		
313	TCGTGTTATCATCATTTCGGAACGCTCATCGCTTTATTGCTATTGGA		
365	AAGCCTGAAATATCTGAGTGCCTCTGAAACCAGCCTCTCGCCTGTGCAGAG		
417	CCGCTGTCAGCAGTTTGTAGCGGTGATCTGGCTGCATGTTCCCTTCGGAA		
469	TATCAGAAATGGCTGGGTACTTACTGATTTAGCCACCATCGCTTATTATCT		
521	ATCAAGAAAAATAACCTCTCTTTAGAGAGGTTTCCCTAGGCCTGA		
573	AGCACCCCTTAGTCTCAATTACCCATAAATTAAAAGGCCTTTTCGTTTA		
625	CTATCATTCAAAAGAGGAAATAGACCAGTTGTCATAAGAATCAGAGTCTAA		
677	TAGAATGAGGTGAAAAGTAAATCACGCAGGATTGTTACTGATAAAGCAGGC		
729	AAGACCTAAAATGTGTTAAGGGCAAAGTGTATTCTTGGCGTCATCCCTTAC	EcoRI	
781	ATATTTGGGTCTTTCTGTAACAAACCTGCCATGAATTGGGAG		

SUBSTITUTE SHEET (RULE 26)

18/28

FIG. 14C

833 GATCGAAACGGCAGATCGCAAAACAGTACATACAGAAGGAGACATGAACAT
 885 GAACATCAAAAAATTGTAAAACAAGCCACAGTACTGACTTTACGACTGCA
 937 CTGCTAGCAGGAGGAGCGACTCAAGCCTCGCGAAAGAAGATATCGATCAAC
 989 GCAATGGTTTATCCAAAGCCTAAAGATGATCCAAGCCAAAGTGTCAACGT
 1041 TTTAGGTGAAGCTCAAAAACCTAATGACTCTCAAGCTCCAAAGCTGATGCG
 1093 CAACAAAATAACTCAACAAAGATCAACAAAGCGCCTCTATGAAATCTTGA
 1145 ACATGCCTAACTTAAACGAAGCGCAACGTAACGGCTTCATTCAAAGTCTTAA
 1197 AGACGACCCAAGCCAAAGCACTAACGTTAGGTGAAGCTAAAAAATTAAAC
 1249 GAATCTCAAGCACCAGAAAGCTGATAACAATTCAACAAAGAACAAACAAATG
 1301 CTTCTATGAAATCTTGAATATGCCACTAAACGAAGAACACGCAATGG
 1353 TTTCATCCAAAGCTTAAAGATGACCCAAGCCAAAGTGCTAACCTATTGTCA
 1405 GAAGCTAAAAAGTTAAATGAATCTCAAGCACCAGCGGATAACAAATTCA
 1457 ACAAAAGAACAAACAAATGCTTCTATGAAATCTTACATTACCTAACTTAA
 1509 CGAAGAACAAACGCAATGGTTCATCCAAAGCCTAAAGATGACCCAAGCCAA
 1561 AGCGCTAACCTTTAGCAGAAGCTAAAAGCTAAATGATGCTCAAGCACCAA
 1613 AAGCTGACAACAAATTCAACAAAGAACAAACAAATGCTTCTATGAAATT
 1665 ACATTACCTAACTTAACGTAAACGTAACGGCTCATCCAAAGCCTT
 1717 AAAGACGATCCGGGAATTCCCGGGATCCGTGACCTGCAGGCATGCAAGC
 1769 TTACTCCCCATCCCTCCAGTAATGACCTCAGAACTCCATCTGGATTGTTC
 1821 AGAACGCTCGGTTGCCGCCGGCGTTTTTATTGGTGAGAACGCAACT
 1873 TGTGCGGCCAATCGAGCCATGTCGTCGTCAACGACCCCCCATTCAAGAACAG
 1925 CAAGCAGCATTGAGAACTTGGAAATCCAGTCCCTCTCCACCTGCTGAGGGC
 1977 AATAAGGGCTGCA CGCGCACTTTATCCGCCTCTGCTGCGCTCCGCCACCGT
 2029 AGTTAAATTATGGTTGGTTATGAAATGCTGGCAGAGACCCAGCGAGACCTG
 2081 ACCGCAGAACAGGCAGCAGGCGTTGCCAGTCAGCGATAACCCGGTTG
 2133 ATAATCAGAAAAGCCCCAAAAACAGGAAGATTGTATAAGCAAATATTAAAT
 2185 TGTAAACGTTAATATTTGTTAAATTCGCGTTAAATTGTTAAATCAGC
 2237 TCATTTTAACCAATAGGCCGAATCGGAAAATCCCTTATAAAATCAAAG
 2289 AATAGCCCGAGATAGGGTTGAGTGTGTTCCAGTTGGAACAAGAGTCCACT
 2341 ATTAAAGAACGTGGACTCCAACGTCAAAGGGCGAAAACCCTATCAGGGC
 2393 GATGGCCCAC TACGTGAACCATCACCCAAATCAAGTTTTGGGGTCGAGGT
 2445 GCCGTAAAGCACTAAATCGGAACCC TAAAGGGAGCCCCGATTAGAGCTT
 2497 ACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAAGAACAGCGAAAGGA
 2549 GCAGGGCGCTAGGGCGCGAGCAAGTGTAGCGGTACGCCCGTAAACCAC
 2601 ACCCGCCCGCTTAATCGCCGCTACAGGGCGGTATCCATTTCGCGAATC
 2653 CGGAGTGTAAAGAAATGAGTCTGAAAGAAAAACACAACTCTGTGTTGCCAAC
 2705 GCATTTGGCTACCCCTGCCACTCACACCATTCAAGGTGCGTCATATAACTGACTG
 2757 AAAACGCCCGCACCGTTGAAGCTGCCAGCGCGCTGGAGCAAGGCACCTGAA
 2809 ACGTATGGCGAGTTGATGGCGGAGTCTCATGCCCTATGCGCGATGATTG
 2861 GAAATCACCGTCCGCAAATTGACACTCTGGTAGAAATGTCAAAGCTGTGA
 2913 TTGGCGACAAGGTGGCGTACGCATGACCGGGCGGGGATTGGCGGCTGTAT
 2965 CGTCGCGCGTATCCCGGAAGAGCTGGTGCCTGCCGCACAGCAAGCTGCGCT
 3017 GAACAAATATGAAGCAAAACAGGTATTAAAGAGACCTTTACGTTGTAAAC

19/28

FIG. 14D

3069 CATCACAGGAGCAGGACAGTGCTGAACGAAACTCCGCACGGCACCCGAT
 3121 GGCAGCCGTACCGACTGTTCTGCCTCGCGTTTGGTATGACGGTGA
 3173 CCTCTGACACATGCAGCTCCGGAGACGGTCACAGCTTGTCTGA
 3225 GCCGGGAGCAGACAAGCCGTAGGGCGCGTCAGCGGGTGTGGCGGGTGT
 3277 GGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCTT
 NdeI
 3329 AACTATGCGGCATCAGAGCAGATTGTA
 3381 AAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGGCTCTCCGC
 3433 TTCCCTCGCTCACTGACTCGCTCGCCTCGGTCGTTGGCTGCGGCAGCGGT
 3485 TCAGCTCACTCAAAGGCGGTAA
 3537 CAGGAAAGAACATGTGAGC
 3589 GCCCGCGTTGCTGGCGTTTCCATAGGCTCCGCC
 3641 AAAAATCGACGCTCAAGTCAGAGGTGGCGAA
 3693 ACCAGGCCTTCCCCCTGGAAGCTCCCTCGCCTCTCCCTCGGG
 3745 GCCGCTTACCGGATACCTGTC
 3797 TCTCATAGCTACGCTGAGGTATCTCAGTT
 3849 AGCTGGGCTGTG
 3901 CGGTAACTATCGTCTTGAGT
 3953 GCAGCAGCCACTG
 4005 CAGAGTTCTGAAGTGG
 4057 TGGTATCTGC
 4109 TCTTGATCCGG
 4161 AGCAGCAGATTACCGC
 4213 TTCTACGGG
 4265 GTCATGAGATT
 4317 GAAGTT
 4369 CCAATGCT
 4421 TCCATAGTT
 4473 TACCATCTGG
 4525 TCCAGATT
 4577 GGT
 4629 CTAGAGTAAGT
 4681 TACAGGC
 4733 GGTT
 4785 CGGTT
 4837 GTT
 4889 TCCG
 4941 AATAGTGT
 4993 TACCG
 5045 TCG
 5097 AACCC
 5149 TTCT
 5201 GCG
 5253 GCAT
 BglII
 4525 TCCAGATT
 4577 GGT
 4629 CTAGAGTAAGT
 4681 TACAGGC
 4733 GGTT
 4785 CGGTT
 4837 GTT
 4889 TCCG
 4941 AATAGTGT
 4993 TACCG
 5045 TCG
 5097 AACCC
 5149 TTCT
 5201 GCG
 5253 GCAT
 Pvul
 4889 TCCG
 4941 AATAGTGT
 4993 TACCG
 5045 TCG
 5097 AACCC
 5149 TTCT
 5201 GCG
 5253 GCAT
 SacI
 4889 TCCG
 4941 AATAGTGT
 4993 TACCG
 5045 TCG
 5097 AACCC
 5149 TTCT
 5201 GCG
 5253 GCAT
 XmnI
 4993 TACCG
 5045 TCG
 5097 AACCC
 5149 TTCT
 5201 GCG
 5253 GCAT

20/28

FIG. 14E

5305 GAAAAATAACAAATAGGGGTTCCGCGCACATTCCCCGAAAAGTGCCACCT
 5357 GACGTCTAAGAAACCATTATTATCATGACATTAAACCTATAAAAATAGGCGTA
 Aval
 5409 TCACGAGGCCCTTCGCTTCAAGCCCGAGGTAAACAAAAAAACAACAGCATA
 5461 AATAACCCCGCTCTTACACATTCCAGCCCTGAAAAGGGCATCAAATTAAAC
 5513 CACACCTATGGTGTATGCATTATTGCATACATTCAATCAATTGTTATCTA
 NdeI
 5565 AGGAAATACTTACATATGGTCGTGCAAACAAACGCAACGAGGCTCTACGAA
 SphI PvuII
 NdeI
 5617 TCGATGCATGCAGCTGATTTCACTTTGCAATTCTACAAACTGCATAACTCA
 5669 TATGTAATCGCTCCTTTAGGTGGCACAAATGTGAGGCATTTCGCTCTT
 5721 TCCGGCAACCACTTCCAAGTAAAGTATAACACACTATACTTATATTCAAA
 5773 AGTGTGTGCTCTGCGAGGCTGTCGGCAGTGGCACAAAACCATAAAACCTT
 5825 TAAGACCTTCTTTTTTACGAGAAAAAGAAACAAAAAACCTGCCCTCT
 5877 GCCACCTCAGCAAAGGGGGTTTGCTCTCGTGTGCTTAAAAATCAGCAA
 5929 GGGACAGGTAGTATTTTGTAGAAGATCACTCAAAAATCTCCACCTTAAA
 5981 CCCTGCCAATTATTATTGTCCGTTGTCTAGCTTACCGAAAGCCAGAC
 6033 TCAGCAAGAATAAAATTATTATTGTCTTCGGTTCTAGTGTAAACGGACAA
 6085 AACCACTCAAAATAAAAGATAACAAGAGAGGGTCTCTCGTATCTTATTCA
 6137 GCAATCGGCCGATTGCTGAACAGATTAATAATAGATTAGCTTTTATT
 6189 TGTGAAAAAAGCTAATCAAATTGTTGTCGGGATCAATTACTGCAAAGTCTC
 6241 GTTCATCCCACCACTGATCTTTAATGATGATTGGGGTGCAAAATGCCAA
 6293 AGGCTTAATATGTTGATATAATTCAATTCCCTCTACTTCATGCGCAA
 6345 CTAGCAGTACCAAGCAATAAACGACTCCGCACCTGTACAAACCGGTGAATCAT
 6397 TACTACGGAGGCCAGCCTCATCACTTGCCTCCATAGATGAATCCGAAC
 6449 CTCATTACACATTAGAACTGCGAATCCATCTCATGGTGAACCAAAGTGA
 6501 CCTAGTTATCGAATAAAACCTATACTCTTTAATATCCCCGACTGGCA
 6553 ATGCCGGGATAGACTGTAACATTCTCACGCATAAAATCCCCCTTCATTTC
 6605 AATGTAATCTATTACCTTATTATTAATTCAATTGCTCATAATTACCTT
 6657 TTTCTTATTACGAAAATGGCCCGATTAAAGCACACCCTTTATTCCGTTAAT
 6709 GCGCCATGACAGCCATGATAATTACTAATACTAGGGAGAAGTTAATAAAC
 6761 TAACCAACATGATTAACATTATTAGAGGTATCGTTCAAAATGGTATGCGT
 6813 TTTGACACATCCACTATATATCCGTGTCGTTCTGTCACCTCTGAATCCC
 6865 TCCAGAAATTCTCTAGCGATTCCAGAAGTTCTCAGAGTCGGAAAGTTGACC
 BglII
 6917 AGACATTACGAACCTGGCACAGATGGTCATAACCTGAAGGAAGATCTGATTG
 6969 TTAAC TGCTTCAGTTAAGACCGAAGCGCTCGTATAACAGATGCGATGAT
 7021 GCAGACCAATCAACATGGCACCTGCCATTGCTACCTGTACAGTCAGGATGG
 7073 TAGAAATGTTGTCGGTCCTTGCACACGAATATTACGCCATTGCGCTGCATAT
 7125 TCAAACAGCTTCTACGATAAGGGCACAAATCGCATCGTGGAACGTTGGG
 7177 CTTCTACCGATTAGCAGTTGATACACTTTCTCTAAGTATCCACCTGAATC
 7229 ATAAATCGGAAAATAGAGAAAAATTGACCATGTGTAAGCGGCCAATCTGAT
 XmnI
 7281 TCCACCTGAGATGCATAATCTAGTAGAATCTCTCGTATCAAATTCACTT
 7333 CCACCTTCACTCACCAGTTGTCATTGATGGCTGAACCTGCTTCCTCTGT
 7385 TGACATGACACACATCATCTCAATATCCGAATAGGGCCATCAGTCTGACGA
 7437 CCAAGAGAGCCATAAACACCAATAGCCTTAACATCATCCCCATATTATCCA
 7489 ATATTGTTCTTAATTGATGAACAATCTCATCTTCTCTAGTCAT
 7541 TATTATTGGTCCATTCACTATTCTCATCCCTTTCAGATAATTAGATT

21/28

FIG. 14F

7593 GCTTTCTAAATAAGAATATTGGAGAGCACC GTTCTTATT CAGCTATTAAT
 7645 AACTCGTCTCCTAAGCATCCTTCAATCCTTTAATAACAATTATAGCATCT
 7697 AATCTTCAACAAACTGGCCGTTGTTGAACTACTCTTAATAAAAATAATTT
 7749 TTCCGTTCCCAATTCCACATTGCAATAATAGAAAATCCATCTCATCGGCTT
 7801 TTTCGTACATCATCTGTATGAATCAAATGCCTTCTCTGTGTCAAGGTT
 7853 TAATTTTTATGTATTCTTTAACAAACCACCATAGGAGATTAACCTTTA
 7905 CGGTGTAACCTCCTCCAAATCAGACAAACGTTCAAATTCTTTCTTCAT
 7957 CATCGGTCAAAATCCGTATCCTTACAGGATATTGCAAGTTGCAAT
 8009 TGCCGATTGTATATCCGATTATATTATTATTTCGGTGCAATCATTGAACT
 8061 TTTACATTGGATCATAGTCTAATTTCATTGCCTTTCCAAAATTGAATCC
 8113 ATTGTTTGATTCACGTAGTTCTGTATTCTTAAATAAGTTGGTCCAC
 8165 ACATACCAATACATGCATGTGCTGATTATAAGAATTATCTTATTATTATT
 8217 GTCACCTCCGTTGCACGCATAAAACCAACAAGATTTTATTAAATTTTTAT
 8269 ATTGCATCATTGGCGAAATCCTTGAGCCATATCTGACAAACTCTTATTAA
 8321 TTCTTCGCCATCATAAACATTAACTGTTAATGTGAGAAACAAACCAACGA
 8373 ACTGTTGGCTTTGTTAAATAACTTCAGCAACAACCTTTGTGACTGAATGC
 8425 CATGTTCATGCTCTCCAGTTGCACATTGGACAAAGCCTGGATTACA
 8477 AAACCACACTCGATACAAACTTTCTTCGCCTGTTCACGATTGTTATAC
 8529 TCTAATATTCAGCACAACTTTACTCTTCAGCCTTTAAATTCAAGAA
 8581 TATGCAGAAGTTCAAAGTAATCAACATTAGCGATTCTTCTCCATGG
 8633 TCTCACTTTCCACTTTGTCTGTCCACTAAAACCCCTGATTTCATCT
 8685 GAATAAAATGCTACTATTAGGACACATAATATTAAAAGAAACCCCCATCTATT
 8737 TAGTTATTGTTAGTCACTTAACTTAAACAGATGGGGTTTCTGTGCA
 8789 ACCAATTTAAGGTTTCAACTTAAACACATACATACCAACACTTCA
 8841 ACGCACCTTCAGCAACTAAAATAAGTACGTTATTCTATATGTATCAA
 Xmn1
 8893 GATAAGAAAGAACAGTCAAAACCATCAAAAAAGACACCTTTCAGGTGC
 8945 TTTTTTATTTATAAAACTCATTCCCTGATCTCGACTCGTTCTTTTAC
 8997 CTCTCGGTTATGAGTTAGTTCAAATTGTTCTTTAGGTTCTAAATCGTGT
 9049 TTTCTTGGATTGTGCTGTTTATCCTTACCTTGTCTACAAACCCCTTAA
 9101 AAACGTTTAAAGGCTTTAAGCCGTGTACGTTCTTAAGG

FIG. 15A

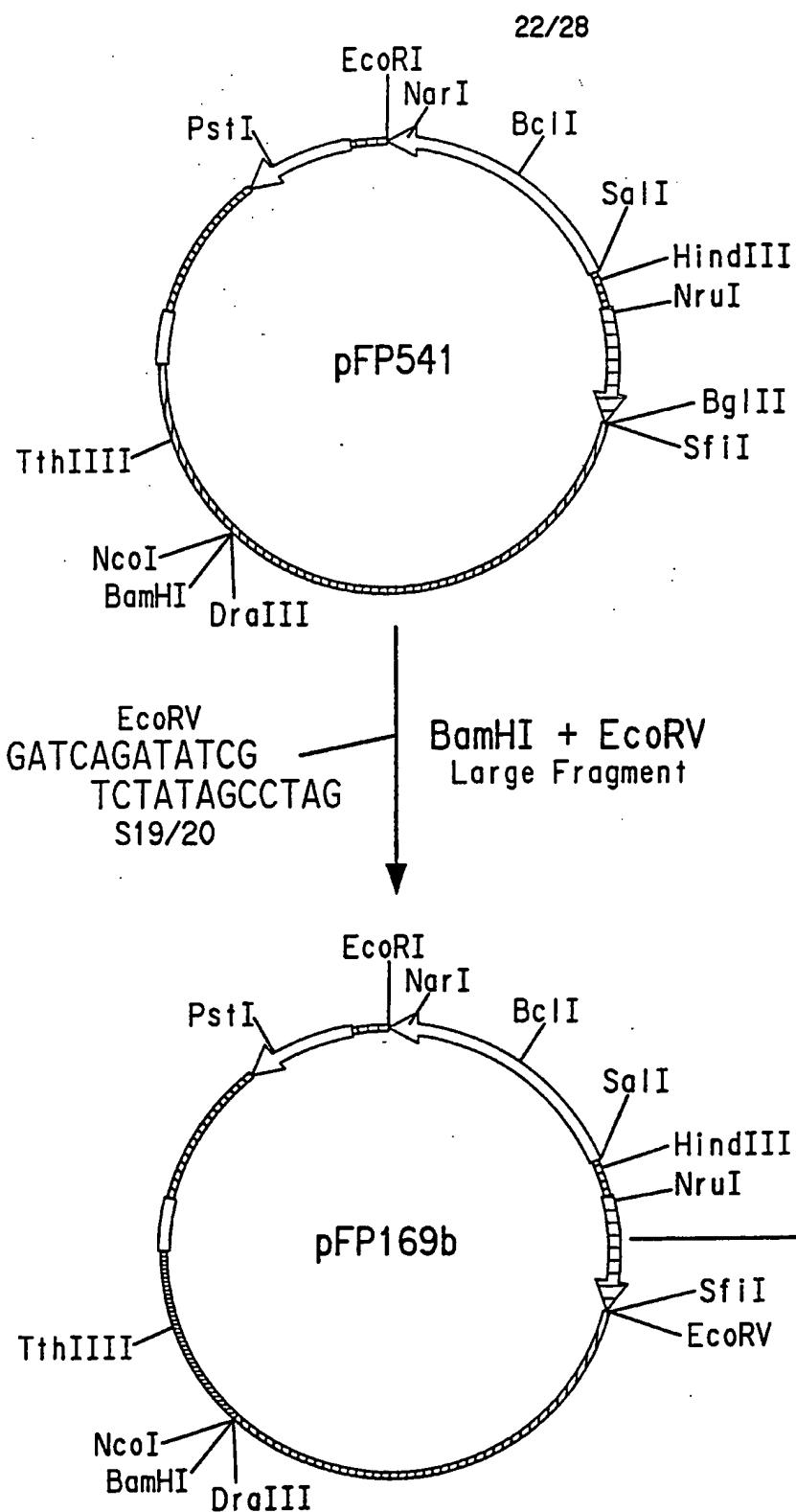
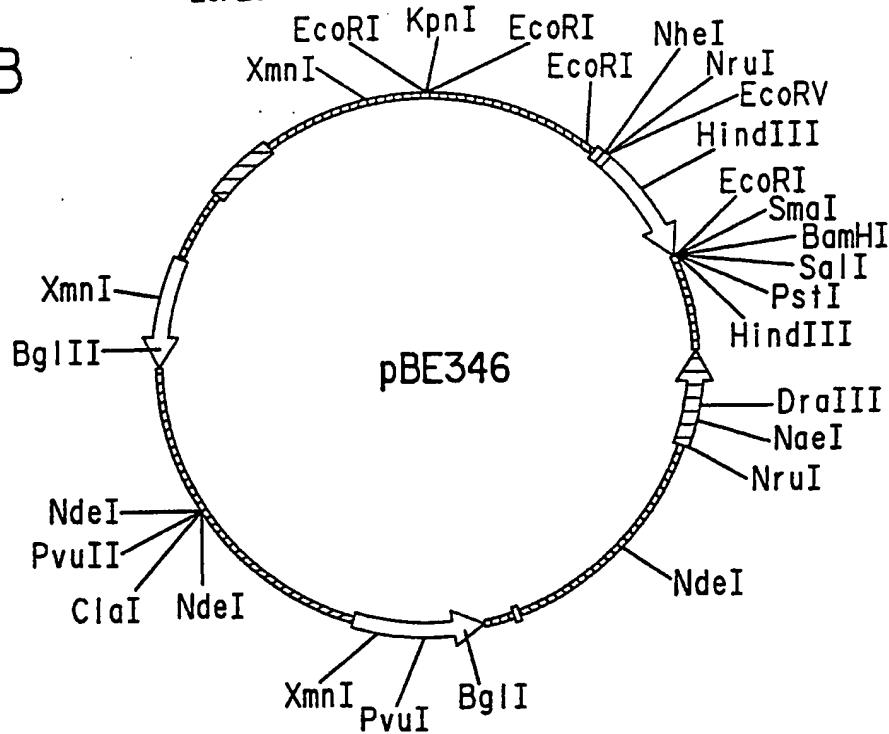


FIG. 15B

23/28



CONTINUED
FROM
FIG. 15A

BamHI + EcoRV

Smaller Fragment

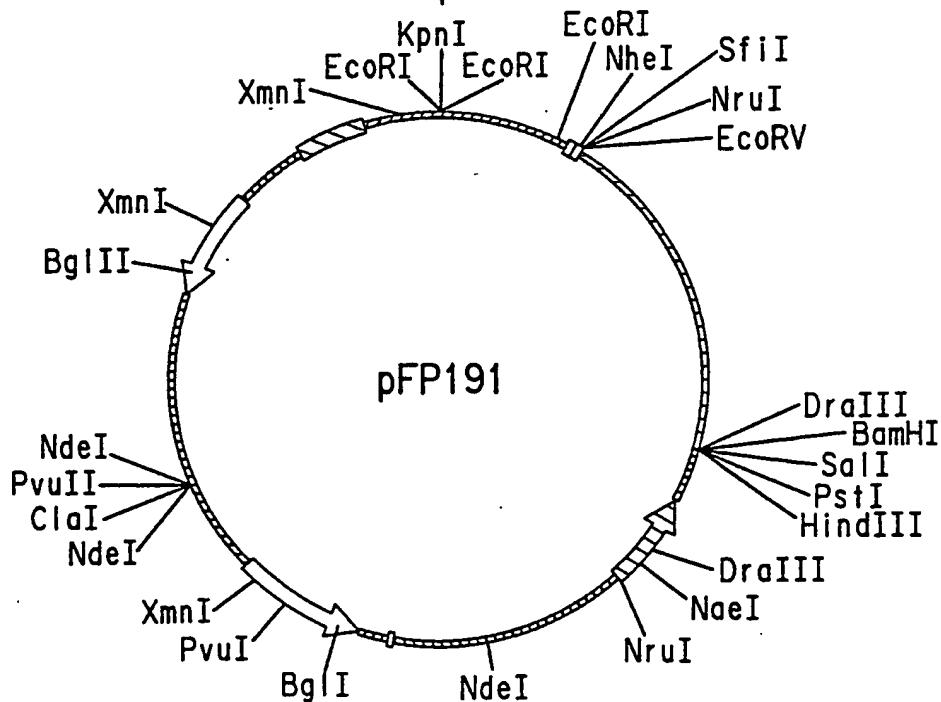
EcoRV + BamHI
Larger Fragment

FIG. 16A

Oligonucleotide P1

GATCTCAAGGAGCCGGTCAAGGGTGGTTACGGAGGTCTGG
 AGTTCTCGGCCAGTTCCACCAATGCTCCAGACCCCTAG
 ▶ Ser Gl nG I aGl yGl nG I y Tyr Gl yLeuGl y

FIG. 16B

Oligonucleotide P2

FIG. 16C

Oligonucleotide P3

FIG. 16D

Oligonucleotide P4

GATCTCAGGGTCTGGTACAGGGGGACAGGGTCCGGGAGCTGCCGCGTGGGGTGGTGGTCAAGGAGGTACGGGTGGTCTTG
 AGTCCCACGACCATCTCCACCTGTCCACGGCCTCGAGGGGGAGGGGACGGACCACTGGTCCCAATGGCCACCGACCTAG
 ➤ Ser Gl nGlyAlaGlyAlaGlyAlaGlyAlaGlyAlaGlyAlaGlyAlaGlyAlaGlyAlaGlyAlaGlyAlaGly

FIG.17

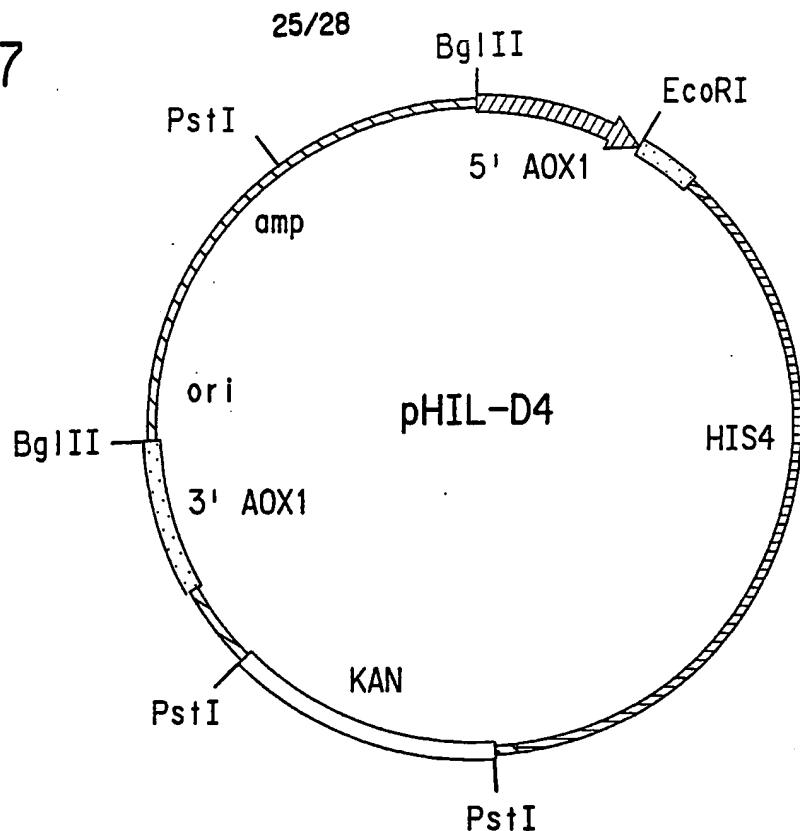
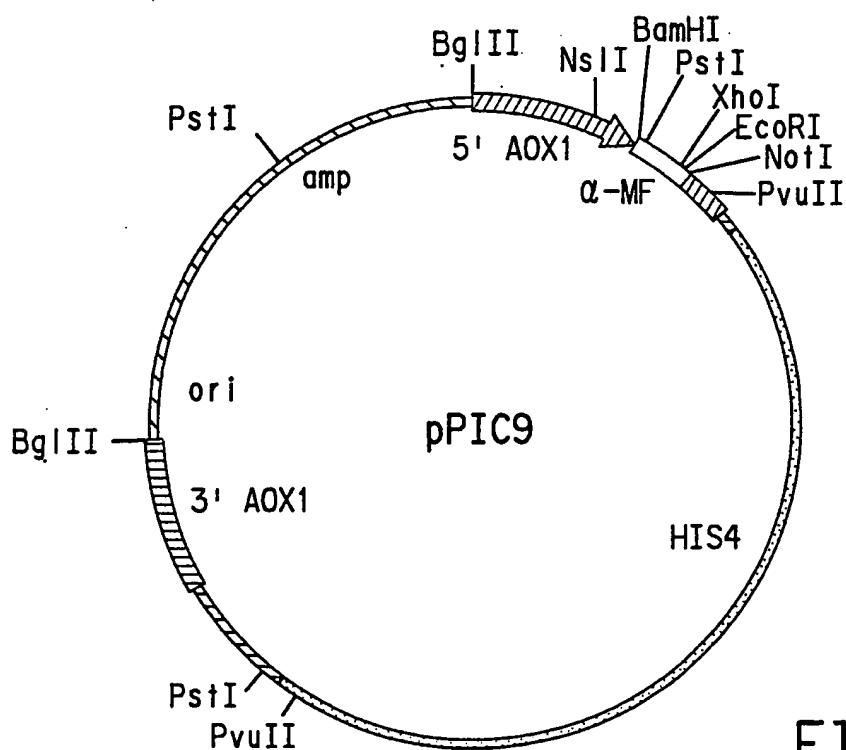


FIG.18



26/28

FIG. 19

NsiI

750 ATGCATTGCTCCACATTGATGCTTCCAAGATTCTGGTGGAAATACTGCTGATA

805 GCCTAACGTTCATGATCAAAATTAACTGTTCTAACCCCTACTTGACAGCAATAT

860 ATAAACAGAAGGAAGCTGCCCTGTCTAACCTTTTTATCATCATTATTAG

915 CTTACTTTCATATTGCGACTGGTCCAATTGACAAGCTTTGATTTAACGACT

970 TTTAACGACAACTTGAGAAGATCAAAAACAACATTATTGAAACGATGAGAT
1► MetArgP1025 TTCCCTCAATTAACTGCAGTTTATTGCAGCATCCTCCGCATTAGCTGCTCC
3► heProSerIlePheThrAlaValLeuPheAlaAlaSerSerAlaLeuAlaAlaPr1080 AGTCAACACTACAACAGAAGATGAAACGGCACAAATTCCGGCTGAAGCTGTCATC
21► oValAsnThrThrGluAspGluThrAlaGlnIleProAlaGluAlaValIle1135 GGTTACTCAGATTAGAAGGGGATTCTCGATGTTGCTGTTGCCATTTCACAA
40► GlyTyrSerAspLeuGluGlyAspPheAspValAlaValLeuProPheSerAsnS1190 GCACAAATAACGGTTATTGTTATAAACTACTATTGCCAGCATTGCTGCTAA
58► erThrAsnAsnGlyLeuLeuPheIleAsnThrThrIleAlaSerIleAlaAlaLy

EcoRI

1245 AGAAGAAGGGGTATCTCTCGAGAAAAGAGAGGGCTGAAGCTTACGTAGAATTCCCT

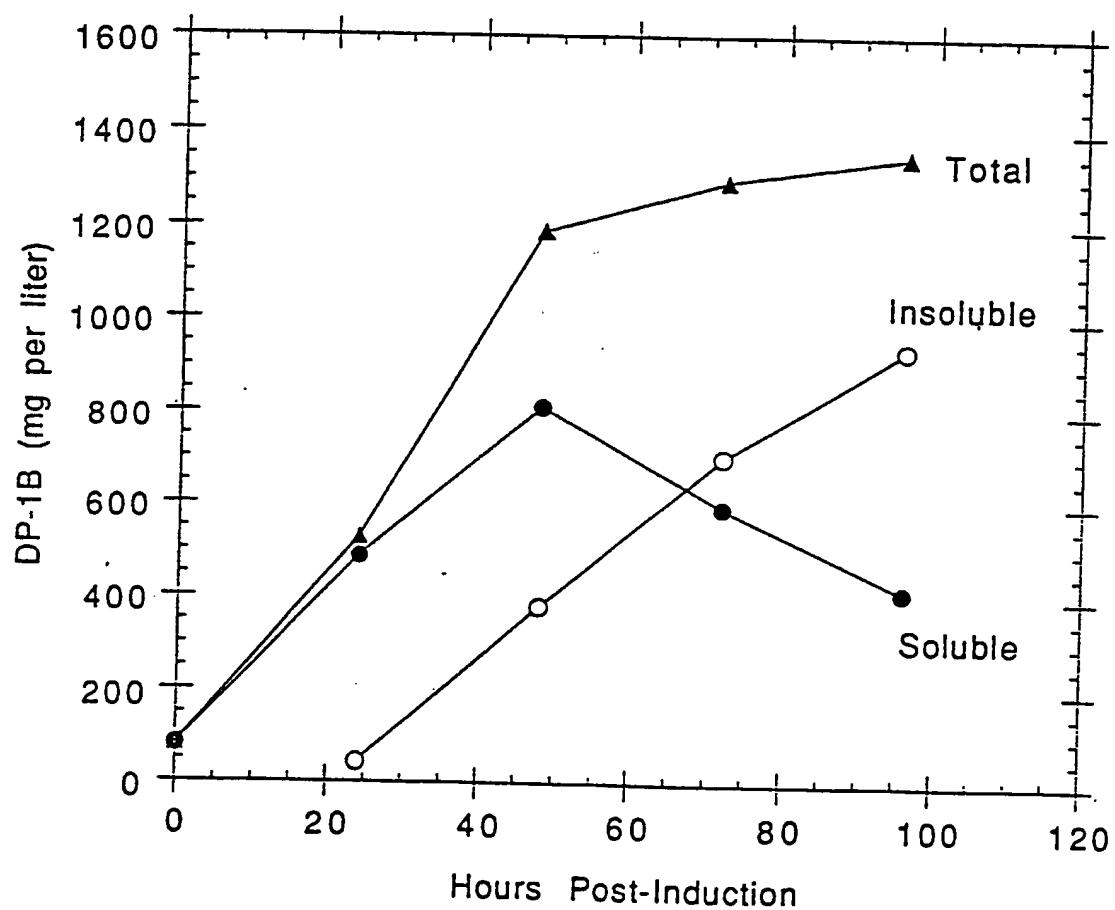
76► sGluGluGlyValSerLeuGluLysArgGluAlaGluAlaTyrValGluPhe SEQ. NO. 97

NotI

1300 AGGGCGGCCGCGAATTAAATTGCCCTAGACATGACTGT SEQ. NO. 96

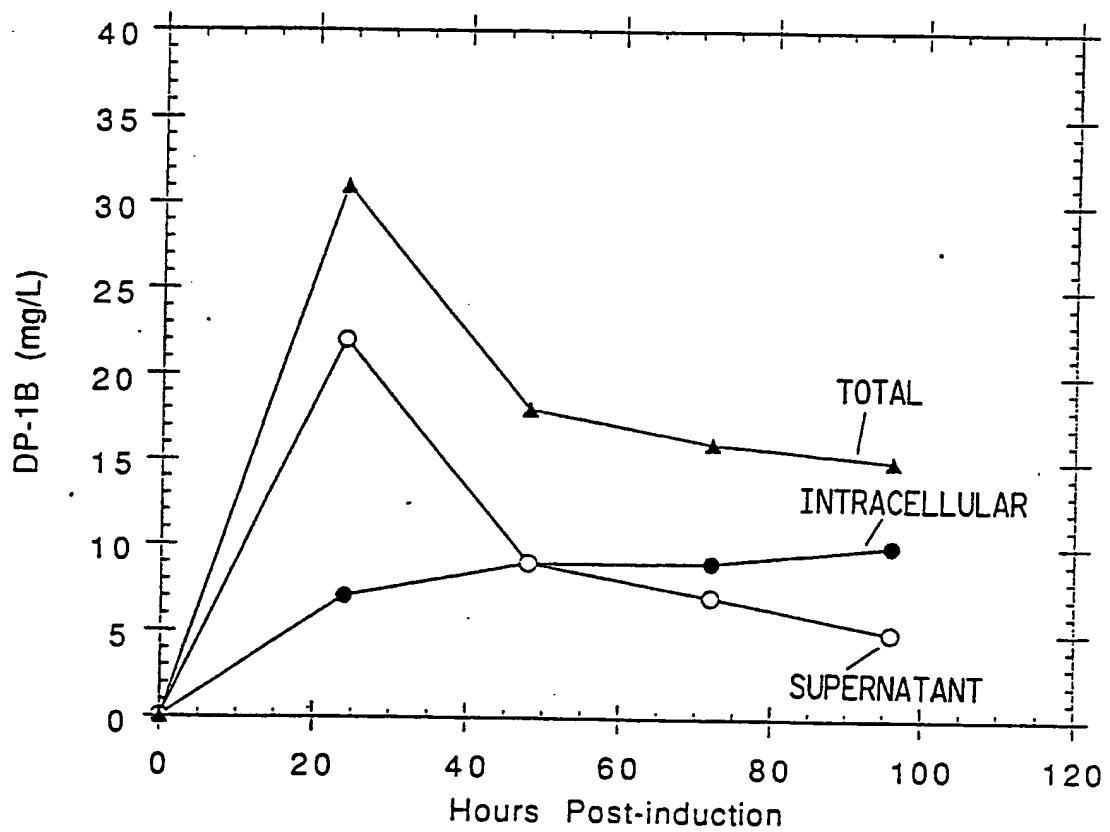
27/28

FIG.20



28/28

FIG.21



MICROORGANISMS

Optional Sheet in connection with the microorganism referred to on page 11, line S 11-17 of the description 1

A. IDENTIFICATION OF DEPOSIT 1

Further deposits are identified on an additional sheet 2

Name of depositary institution 4

AMERICAN TYPE CULTURE COLLECTION

Address of depositary institution (including postal code and country) 4

12301 Parklawn Drive
Rockville, Maryland 20852
US

Date of deposit 5

15 June 1993 (15.06.93)

Accession Number 6

ATCC 69328

B. ADDITIONAL INDICATIONS 7 (leave blank if not applicable). This information is continued on a separate attached sheet

In respect of those designations in which a European patent is sought, a sample of the deposited microorganism will be made available until the publication of the mention of the grant of the European patent or until the date on which the application has been refused or withdrawn or is deemed to be withdrawn, only by the issue of such a sample to an expert nominated by the person requesting the sample. (Rule 28(4) EPC)

C. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE 8 (If the indications are not for all designated States)

D. SEPARATE FURNISHING OF INDICATIONS 9 (leave blank if not applicable)

The indications listed below will be submitted to the International Bureau later 9 (Specify the general nature of the indications e.g., "Accession Number of Deposit")

E. This sheet was received with the international application when filed (to be checked by the receiving Office)G. Aaron Smith
PCT International Division

(Authorized Officer)

 The date of receipt (from the applicant) by the International Bureau 10

wsb

(Authorized Officer)

MICROORGANISMS

Optional Sheet in connection with the microorganism referred to on page 11, line S 11-17 of the description.

A. IDENTIFICATION OF DEPOSIT¹

Further deposits are identified on an additional sheet.

Name of depositary institution⁴

AMERICAN TYPE CULTURE COLLECTION

Address of depositary institution (including postal code and country)⁴12301 Parklawn Drive
Rockville, Maryland 20852
USDate of deposit⁵

15 June 1993 (15.06.93)

Accession Number⁶

ATCC 69327

B. ADDITIONAL INDICATIONS⁷ (leave blank if not applicable). This information is continued on a separate attached sheet

In respect of those designations in which a European patent is sought, a sample of the deposited microorganism will be made available until the publication of the mention of the grant of the European patent or until the date on which the application has been refused or withdrawn or is deemed to be withdrawn, only by the issue of such a sample to an expert nominated by the person requesting the sample. (Rule 28(4) EPC)

C. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE⁸ (if the indications are not for all designated States)D. SEPARATE FURNISHING OF INDICATIONS⁹ (leave blank if not applicable)

The indications listed below will be submitted to the International Bureau later¹⁰ (Specify the general nature of the indications e.g. "Accession Number of Deposit")

E. This sheet was received with the international application when filed (to be checked by the receiving Office)

G. Anton Smith
PCT International Division

(Authorized Officer)

 The date of receipt (from the applicant) by the International Bureau is

was

(Authorized Officer)

MICROORGANISMS

Optional Sheet in connection with the microorganism referred to on page 11, line 5, 11-17 of the description.

A. IDENTIFICATION OF DEPOSIT

Further deposits are identified on an additional sheet.

Name of depositary institution

AMERICAN TYPE CULTURE COLLECTION

Address of depositary institution (including postal code and country)

12301 Parklawn Drive
Rockville, Maryland 20852
US

Date of deposit

15 June 1993 (15.06.93)

Accession Number

ATCC 69326

B. ADDITIONAL INDICATIONS (leave blank if not applicable). This information is continued on a separate attached sheet

In respect of those designations in which a European patent is sought, a sample of the deposited microorganism will be made available until the publication of the mention of the grant of the European patent or until the date on which the application has been refused or withdrawn or is deemed to be withdrawn, only by the issue of such a sample to an expert nominated by the person requesting the sample. (Rule 28(4) EPC)

C. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE (If the indications are not for all designated States)

D. SEPARATE FURNISHING OF INDICATIONS (leave blank if not applicable)

The indications listed below will be submitted to the International Bureau later. (Specify the general nature of the indications e.g., "Accession Number of Deposit")

E. This sheet was received with the international application when filed (to be checked by the receiving Office)G. Anton Smith
PCT International Division

(Authorized Officer)

 The date of receipt (from the applicant) by the International Bureau

was

(Authorized Officer)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- BLACK BORDERS**
- IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- FADED TEXT OR DRAWING**
- BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- SKEWED/SLANTED IMAGES**
- COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- GRAY SCALE DOCUMENTS**
- LINES OR MARKS ON ORIGINAL DOCUMENT**
- REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.